

# Contents

<b>Publications out of this thesis</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>Table of contents</b>	<b>xii</b>
<b>List of figures</b>	<b>xiv</b>
<b>List of tables</b>	<b>xv</b>
<b>Acronyms</b>	<b>xvii</b>
<b>1 Overview of this thesis</b>	<b>1</b>
1.1 Overview of the chapter . . . . .	1
1.2 The context and justification for this research . . . . .	1
1.3 Statement of the problem . . . . .	5
1.4 Background . . . . .	11
1.5 Summary and looking ahead . . . . .	15
<b>2 Spatial data infrastructures and geospatial data</b>	<b>19</b>
2.1 Overview of the chapter . . . . .	19
2.2 Spatial data infrastructure . . . . .	20
2.3 Geospatial data or geographical information . . . . .	30
2.4 Classification, taxonomy, ontologies, folksonomies, etc . . . . .	39
2.5 Models for geospatial data in a GIS . . . . .	48
2.6 Formal models . . . . .	51
2.7 Data quality and metadata . . . . .	53
2.8 Incremental updating and versioning . . . . .	54
2.9 Cartography . . . . .	55
2.10 Virtual globes and geobrowsers . . . . .	57
2.11 Summary and looking ahead . . . . .	58
<b>3 The context for user-generated content</b>	<b>59</b>
3.1 Overview of the chapter . . . . .	59
3.2 Inter-networking technologies . . . . .	61
3.3 The dot.com bubble . . . . .	67

## CONTENTS

## CONTENTS

3.4	After the dot.com bust . . . . .	68
3.5	The Semantic Web . . . . .	77
3.6	Social media services . . . . .	79
3.7	Social mapping . . . . .	83
3.8	Controlling the Internet . . . . .	83
3.9	Privacy, censorship and liability . . . . .	94
3.10	The right to exploit content . . . . .	104
3.11	Curation . . . . .	108
3.12	The digital divide . . . . .	110
3.13	Standards . . . . .	115
3.14	Summary and looking ahead . . . . .	116
<b>4</b>	<b>User-generated content and volunteered geographical information</b>	<b>119</b>
4.1	Overview of the chapter . . . . .	119
4.2	User-generated content, crowd-sourcing, citizen science and neogeography are not the same! . . . . .	121
4.3	User-generated content . . . . .	123
4.4	Citizen science . . . . .	136
4.5	Volunteered geographical information . . . . .	142
4.6	Crowd source . . . . .	150
4.7	Neogeography . . . . .	157
4.8	Validity of user-generated content in scholarly research . . . . .	160
4.9	Citing . . . . .	175
4.10	Summary and looking ahead . . . . .	178
<b>5</b>	<b>Metadata</b>	<b>179</b>
5.1	Overview of the chapter . . . . .	179
5.2	Definition of metadata . . . . .	181
5.3	Aspects of metadata . . . . .	183
5.4	Encoding metadata . . . . .	184
5.5	Metadata and specifications . . . . .	186
5.6	Metadata tools . . . . .	188
5.7	Categories of metadata . . . . .	189
5.8	Standards for metadata . . . . .	194
5.9	The limitations of metadata . . . . .	201
5.10	Metadata vs searching . . . . .	203
5.11	Metadata and linked open data . . . . .	204
5.12	VGI and metadata . . . . .	205
5.13	Summary and looking ahead . . . . .	206
<b>6</b>	<b>Quality</b>	<b>209</b>
6.1	Overview of the chapter . . . . .	209
6.2	Aspects of geospatial data quality . . . . .	211
6.3	Four stages of recognising the quality of a resource . . . . .	213
6.4	GNSS errors . . . . .	216
6.5	Commonly used dimensions of quality for geospatial data . . . . .	218

## CONTENTS

## CONTENTS

6.6	Further perspectives on the dimensions of quality . . . . .	225
6.7	Quality of volunteered geographical information . . . . .	230
6.8	Assessing the quality of several VGI repositories . . . . .	235
6.9	Using quality to classify geospatial data . . . . .	246
6.10	Standards for the quality of geospatial data . . . . .	248
6.11	Summary and looking ahead . . . . .	249
<b>7</b>	<b>Perceptions of virtual globes, VGI and SDIs</b>	<b>251</b>
7.1	Overview of the chapter . . . . .	251
7.2	Background to the questionnaire . . . . .	252
7.3	Summary of the results from CODIST-I . . . . .	254
7.4	Summary of the results from GISSA Gauteng . . . . .	261
7.5	Analysis of the results from CODIST-I and GISSA . . . . .	267
7.6	Conclusions . . . . .	269
7.7	Summary and looking ahead . . . . .	269
<b>8</b>	<b>Assessing qualitatively taxonomies of user generated content</b>	<b>271</b>
8.1	Overview of the chapter . . . . .	271
8.2	Taxonomies of UGC and VGI . . . . .	272
8.3	Repositories containing VGI used for assessing taxonomies . . . . .	274
8.4	Qualitative assessment of published taxonomies of UGC . . . . .	293
8.5	VGI repositories and citizen science . . . . .	310
8.6	Summary of the qualitative assessment . . . . .	310
8.7	Preliminary taxonomy of user generated content . . . . .	311
8.8	Summary and looking ahead . . . . .	314
<b>9</b>	<b>Using formal concept analysis to assess taxonomies</b>	<b>315</b>
9.1	Overview of the chapter . . . . .	315
9.2	Formal concept analysis (FCA) . . . . .	317
9.3	FCA and the feature model . . . . .	329
9.4	Applying formal concept analysis to assess taxonomies . . . . .	330
9.5	Stability exploration . . . . .	338
9.6	Assessing the discrimination adequacy of existing taxonomies of UGC . . . . .	341
9.7	Summary and looking ahead . . . . .	352
<b>10</b>	<b>Conclusions</b>	<b>355</b>
10.1	Overview . . . . .	355
10.2	Review . . . . .	356
10.3	Future research topics . . . . .	358
<b>A</b>	<b>Questionnaire on VGI</b>	<b>361</b>
<b>B</b>	<b>Published taxonomies of user generated content</b>	<b>365</b>
B.1	Overview of the appendix . . . . .	365
B.2	OECD Working Party on the Information Economy . . . . .	365
B.3	Gervais' taxonomy for copyright issues . . . . .	368
B.4	Budhathoki, Nedovic-Budic and Bruce's framework for VGI . . . . .	369

## CONTENTS

## CONTENTS

B.5 Coleman, Georgiadou and Labonte's nature and motivation of producers .	371
B.6 Castelein, Grus, Crompvoets and Bregt's characterization of repositories of VGI . . . . .	373
<b>Bibliography</b>	<b>375</b>
<b>Web pages</b>	<b>451</b>
<b>Colophon</b>	<b>459</b>

# List of Figures

1.1	Examples of online repositories of user-generated content . . . . .	3
1.2	Geospatial content for an SDI can be official, commercial and/or VGI . . .	5
1.3	The value chain from content to SDI . . . . .	9
1.4	Overview of how the chapters link together . . . . .	16
2.1	The high-level UML classes of the enterprise viewpoint of an SDI [Hjel- mager <i>et al</i> 2008]. . . . .	21
2.2	A feature and related concepts . . . . .	35
3.1	Jack Warner seemingly holding The Onion that fooled him [Topping 2015].	66
4.1	Peckham’s peace wall, a response to the riots in the UK in August 2011 (photo from BBC [2016]) . . . . .	125
4.2	<i>Wikipedia</i> , as commented on by the cartoon strip <i>Alex</i> [Peattie, Charles and Taylor, Russell 2016], from 4 June 2008. . . . .	134
4.3	OpenStreetMap data of Port Alfred, as at 1 February 2012. . . . .	148
6.1	Stages of recognising data quality . . . . .	214
6.2	Conceptual model of quality for geographic data [ISO 19157 2013]. . . . .	216
6.3	GPS positional accuracy issues . . . . .	216
6.4	Dimensions of data quality . . . . .	219
6.5	Alleged pirate boats on the beach at Eyl, Somalia [“expedition” 2009]. . . .	234
6.6	SABAP2 distribution of Southern Black Korhaan, as at 24 June 2012 [Ani- mal Demography Unit 2016 <i>b</i> ]. . . . .	237
6.7	OpenStreetMap showing the Gauteng area [OpenStreetMap 2016]. . . . .	240
6.8	Tracks4Africa showing the Ponto Do Ouro area [Tracks4Africa 2016]. . . .	243
6.9	Types of VGI from the perspective of quality [Cooper <i>et al</i> 2011 <i>a</i> ]. . . . .	247
8.1	Tracks4Africa, showing the area around Nylsvley . . . . .	277
8.2	OpenStreetMap, showing the area around Nylsvley . . . . .	278
8.3	Wikimapia, showing the area around Nylsvley . . . . .	279
8.4	Bing Maps, showing the area around Nylsvley . . . . .	280
8.5	Google Earth . . . . .	281
8.6	Google Maps . . . . .	282
8.7	2nd South African Bird Atlas Project (SABAP2) . . . . .	283
8.8	NAVTEQ’s <i>Map Reporter</i> . . . . .	284

*LIST OF FIGURES*

*LIST OF FIGURES*

8.9	Tom Tom's <i>Map Share</i> . . . . .	285
8.10	Mobilitate . . . . .	286
8.11	HarassMap . . . . .	287
8.12	OpenAddresses, showing address data donated by Paarl . . . . .	290
8.13	NaturalWorld . . . . .	291
8.14	Examples of VGI used to create virtual land art . . . . .	302
9.1	An example of a line diagram . . . . .	318
9.2	The lattice shown in Figure 9.1, but with reduced labelling and a different layout . . . . .	319
9.3	A very stable, but rather boring, lattice. . . . .	322
9.4	The girls of Ipanema [Watson <i>et al</i> 2012], adapted from [Jobim <i>et al</i> 1962/1964].	324
9.5	A lattice used to show lower bounds for stability indices. . . . .	326
9.6	Attribute exploration in ConExp [Yevtushenko <i>et al</i> 2003]. . . . .	328
9.7	Feature model in FCA . . . . .	330
9.8	Feature instances in FCA . . . . .	331
9.9	Absent and redundant attributes and objects. . . . .	333
9.10	The taxonomy of Gervais [2009] for copyright issues for UGC. . . . .	335
9.11	A subset of the taxonomy of Coleman <i>et al</i> [2009] for assessing the nature and motivation of <i>producers</i> . . . . .	336
9.12	Figure 9.11 with only the repositories in Coleman <i>et al</i> [2009]. . . . .	337
9.13	OECD's social drivers of UGC [Wunsch-Vincent & Vickery 2007]. . . . .	342
9.14	Copyright issues, from Gervais [2009]. . . . .	343
9.15	The ten taxonomies discriminated by only five attributes from Budhathoki <i>et al</i> [2010]. . . . .	345
9.16	Expertise of producers [Coleman <i>et al</i> 2009]. . . . .	346
9.17	Expertise of producers and contexts for contributing [Coleman <i>et al</i> 2009].	347
9.18	Characterization of VGI repositories [Castelein <i>et al</i> 2010] . . . . .	349
9.19	The ten taxonomies discriminated by only five attributes from Castelein <i>et al</i> [2010]. . . . .	350
9.20	Wiggins and Crowston's typology of citizen science and VGI repositories .	351

# List of Tables

1	Acronyms . . . . .	xvii
2.1	Features and related concepts . . . . .	36
2.2	ISCO-88 (the old version) vs ISCO-08 (the new version) [International Labour Organization 2016] . . . . .	46
2.3	Real and virtual maps [Moellering 2000] . . . . .	56
4.1	VGI, crowd-sourcing, citizen science & neogeography [Cooper 2015]. . . . .	122
5.1	Encoded and free-text metadata . . . . .	185
6.1	Three GPS records in sequence showing a gross error . . . . .	218
6.2	VGI and the dimensions of quality . . . . .	231
7.1	Reasons for using a virtual globe/geobrowser (CODIST respondents) . . . . .	259
7.2	Reasons for using a virtual globe/geobrowser (GISSA respondents) . . . . .	265
8.1	Rankings of repositories as at 2 April 2014, according to Alexa [2014] . . . . .	275
8.2	Characteristics of repositories containing VGI . . . . .	288
8.3	VGI repositories and UCC drivers . . . . .	294
8.4	VGI repositories and copyright issues . . . . .	299
8.5	VGI repositories and VGI framework . . . . .	302
8.6	VGI repositories and producers' motivation . . . . .	305
8.7	VGI repositories and repository characterization . . . . .	308
8.8	VGI repositories and citizen science . . . . .	310
9.1	A cross-table of the formal context shown in Figures 9.1 and 9.2. . . . .	320
9.2	A many-valued context and the equivalent one-valued context. . . . .	320
9.3	The subsets of the concept covering {Alice, Amy}, with their derivatives. . . . .	324
A.1	Questionnaire on virtual globes and geobrowsers . . . . .	361

*LIST OF TABLES*

*LIST OF TABLES*

---

# Acronyms

Table 1: Acronyms

Acronym	Expansion
ABCD	asset-based community development
ADU	Animal Demography Unit
AGILE	Association of Geographic Information Laboratories for Europe
AJAX	asynchronous JavaScript and XML
AM/FM	automated mapping/facilities management
ANSI	American National Standards Institute
API	application program interface
ARPA	Advanced Research Projects Agency
ARPANET	Advanced Research Projects Agency Network
ASCII	American Standard Code for Information Interchange
ASM	abstract state machine
ASN.1	abstract syntax notation one
ASP	application service provider
B2B	business to business
BBC	British Broadcasting Corporation
BBS	bulletin board system
BIRP	bird in reserves project
BPA	bloodstain pattern analysis
BSI	British Standards Institute
C2C	citizen to citizen
CAD	computer-aided design
CADP	construction and analysis of distributed processes
CASL	common algebraic specification language
CCITT	Comité Consultatif International Téléphonique et Télégraphique (International Telegraph and Telephone Consultative Committee)
CD	Committee Draft
CD-ROM	Compact Disc Read Only Memory
CDSM	Chief Directorate: Surveys and Mapping
CEN	Comité Européen de Normalisation (European Committee for Standardization)
CLA	cultural and linguistic adaptability

*Continued on next page*



## Acronyms

Acronym	Expansion
CMS	content management system
CODI	Committee for Development Information
CODATA	Committee on Data for Science and Technology (of ICSU)
CODIST	Committee for Development Information, Science and Technology
CSDGM	content standard for digital geospatial metadata
CSI	Committee for Spatial Information
CSIR	Council for Scientific and Industrial Research
CSP	communicating sequential processes
CSS	cascading style sheets
CWAC	co-ordinated waterbird counts
DARPA	Defense Advanced Research Projects Agency
DDI	Data Documentation Initiative
DDoS	distributed denial of service attack
DIN	Deutsches Institut für Normung
DIS	Draft International Standard
DOAP	description of a project
DOI	digital object identifier
DQAF	Data Quality Assessment Framework
DRDLR	Department of Rural Development and Land Reform
DTD	document type definition
DVD	digital versatile disc
EFF	Electronic Frontier Foundation
ETL	extract, transfer, load
EULA	end-user licence agreement
EUOSME	European Open Source Metadata Editor
FBI	Federal Bureau of Investigation
FCA	formal concept analysis
FDIS	Final Draft International Standard
FGDC	Federal Geographic Data Committee
FOAF	friend of a friend
FTP	file transfer protocol
FTC	Federal Trade Commission
GIS	geographical information system
GISc	geographical information science
GISSA	Geo-Information Society of South Africa
GLONASS	global navigation satellite system
GML	geography markup language
GNSS	global navigation satellite systems
GNU	GNU not unix
GPS	global positioning system
HTML	hypertext markup language
HTTP	hypertext transfer protocol
IANA	Internet Assigned Numbers Authority
ICA	International Cartographic Association

*Continued on next page*

## Acronyms

Acronym	Expansion
ICSTI	International Council for Scientific and Technical Information
ICT	information and communications technology
ICSU	International Council for Science
IDEF	integration definition methods
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IMF	International Monetary Fund
INCITS	International Committee for Information Technology Standards
INSPIRE	Infrastructure for Spatial Information in the European Community
IP	internet protocol
IS	International Standard
ISO	International organization for Standardization
ISP	Internet service provider
ISPRS	International Society of Photogrammetry and Remote Sensing
IT	information technology
ITU	International Telecommunication Union
ITU-T	Telecommunication Standardization Sector ( <i>of ITU</i> )
JPEG	Joint Photographic Experts Group
JRC	Joint Research Centre
JSON	JavaScript object notation
JTC	Joint Technical Committee ( <i>of ISO and IEC</i> )
KML	keyhole markup language
LBS	location-based service
LIS	land information system
LOD	linked open data
MAfA	Mapping Africa for Africa
MARC	machine-readable cataloging
MAUP	mappable areal unit problem
MIME	multipurpose internet mail extensions
MMOG	massively multiplayer online game
MOF	meta-object facility
MP3	MPEG-1 audio layer 3
MPEG	Moving Picture Experts Group
MRDB	multiple representation database
MSC	message sequence chart
NAP	North American Profile ( <i>of ISO 19115</i> )
NASA	National Aeronautics and Space Administration
NCDCDS	National Committee for Digital Cartographic Data Standards
NCP	network control protocol
NDP	National Development Plan
NGI	National Geo-spatial Information (Chief Directorate)
NGO	non-governmental organisation
NMA	national mapping agency

*Continued on next page*



## Acronyms

Acronym	Expansion
NOAA	National Oceanic and Atmospheric Administration
NSIF	National Spatial Information Framework
NWIP	new work item proposal
OASIS	Organization for the Advancement of Structured Information Standards
OECD	Organisation for Economic Co-operation and Development
OGC	Open Geospatial Consortium, Inc
OMG	Object Management Group
OWL	web ontology language
P2P	peer to peer
PAIA	Promotion of Access to Information Act (Act No 2 of 2000)
PDA	personal digital assistant
PHP	personal home page
PKI	public key infrastructure
PPGIS	public participatory geographical information system
PSI	published subject indicator
RDF	resource description framework
RFC	request for comments
RFID	radio frequency identification
RM ODP	reference model for open distributed processing
RPG	role playing game
RSS	really simple syndication
SaaS	software as a service
SAB	service area business
SABAP1	First Southern African Bird Atlas Project
SABAP2	Second Southern African Bird Atlas Project
SABCA	Southern African Butterfly Conservation Assessment
SABS	South African Bureau of Standards
SAC	Satellite Applications Centre
SADC	Southern African Development Community
SADCA	SADC Cooperation in Accreditation
SADCMEL	SADC Cooperation in Legal Metrology
SADCMET	SADC Cooperation in Measurement Traceability
SADCO	Southern African Data Centre for Oceanography
SADCSTAN	Southern African Development Community Cooperation in Standardization
SAEON	South African Environmental Observation Network
SAGIMS	South African Geo-spatial Information Management Strategy
SANSA	South African National Space Agency
SARCA	Southern African Reptile Conservation Assessment
SASDI	South African Spatial Data Infrastructure
SASQAF	South African Statistical Quality Assessment Framework
SDI	spatial data infrastructure
SDMX	statistical data and metadata exchange

*Continued on next page*

## Acronyms

Acronym	Expansion
SMS	short message service
SNA	systems network architecture
SNS	social networking site
SOA	service-oriented architecture
SOAP	simple object access protocol
SPARQL	SPARQL protocol and RDF query language
SPEM	software and systems process engineering meta-model specification
SPISYS	Spatial Planning Information System
SPLUMA	Spatial Planning and Land Use Management Act (Act No 16 of 2013)
SQL	structured query language
SVG	scalable vector graphics
SWGSTAIN	Scientific Working Group on Bloodstain Pattern Analysis (established by the FBI)
SWE	sensor web enablement
TCP	transmission control protocol
TM	topic map
TS	Technical Specification
UCC	user created content
UGC	user generated content
UDDI	universal description, discovery and integration
UML	unified modelling language
UN	United Nations
UN ECA	United Nations Economic Commission for Africa
UN ECE	United Nations Economic Commission for Europe
UNIDO	United Nations Industrial Development Organization
UNOOSA	United Nations Office for Outer Space Affairs
UNSD	United Nations Statistical Division
UPU	Universal Postal Union
URI	uniform resource identifier
URL	universal resource locator
URN	uniform resource name
VGI	volunteered geographical information
VDM	Vienna development method
VoID	Vocabulary Of Interlinked Datasets
VOIP	voice over internet protocol
VOS	Voluntary Observing Ships
VPN	virtual private network
VV&A	verification, validation and accreditation
VV&C	verification, validation and certification
WAIS	wide area information servers
WD	Working Draft
WFS	web feature service
WG	working group
WGS	world geodetic system

*Continued on next page*

## Acronyms

Acronym	Expansion
WIPO	World Intellectual Property Organization
WMO	World Meteorological Organization
WMS	web map service
WSSN	World Standard Services Network
WTO	World Trade Organization
WWW	world wide web
W3C	World Wide Web Consortium
XBRL	eXtensible Business Reporting Language
XHTML	extensible hypertext markup language
XMI	XML Metadata Interchange
XML	extensible markup language
XMPP	extensible messaging and presence protocol
XSD	XML schema document
XSLT	extensible stylesheet language transformations

---

## Chapter 1

# Overview of this thesis

### 1.1 Overview of the chapter

This thesis presents an analysis of the nature of *volunteered geographical information* (VGI) and on its applicability for use in a *spatial data infrastructure* (SDI) to supplement official and commercial sources, particularly given the ease with which ordinary people can document their environment, experiences, perspectives and prejudices, share them widely and rapidly, and even query anyone else's content. For this research, taxonomies and repositories of such information were examined qualitatively and using *formal concept analysis* (FCA). Further, this thesis attempts to reflect on the context for SDIs and VGI and the challenges and opportunities for both.

This chapter introduces the context for this thesis. Section 1.2 provides an overview of the key concepts: geospatial data, SDIs, inter-networking, user-generated content (UGC), VGI, classification, folksonomies, citizen science, crowd sourcing, neogeography, meta-data, quality, standards and FCA. Section 1.3 provides the statement of the problem being addressed in this thesis: what the problem is, why it is important, what I did in researching the problem and the contribution. Section 1.4 provides some background for the rest of the thesis, covering the literature reviewed and used, the terminology and the limitations of the Internet. Section 1.5 introduces all the other chapters and appendices.

### 1.2 The context and justification for this research

The word *geospatial* is defined as “*relating to or denoting data that is associated with a particular location*” [Oxford Dictionaries 2016]. One of the distinguishing characteristics of the use of *geospatial data* is that the same, common, base data sets are used by many different users for many diverse applications. Hence, there is a growing need to share and organise geospatial data across different disciplines and organisations, which has resulted in the development and implementation of *spatial data infrastructures* (SDIs) and of the theory and notions behind them, see Section 2.2. An SDI is an evolving concept about

---

## 1. Overview of this thesis

---

facilitating and coordinating the exchange and sharing of geospatial data and services between stakeholders from different levels in the geospatial data community [Hjelmager *et al* 2008]. An SDI is more than just the technology of a distributed *geographical information system* (GIS — see Section 2.5): it is generally considered to be the collection of technologies, policies and institutional arrangements that facilitates the availability of, and access to, geospatial data. It provides a basis for geospatial data discovery, evaluation and application for a variety of users and providers [Nebert 2004].

The Internet has spawned the development of *virtual communities* or *virtual social networks* which share data with one another, and with the public at large. This *user generated content* (UGC, see Section 4.3) is most obvious in Web sites such as Wikipedia [Wikimedia 2016], the free, online encyclopaedia in many languages, consisting of contributions mainly from the public at large, rather than from domain experts (though it does also include much content from encyclopaedias that are out of copyright and from other expert sources). Similarly, virtual communities have also facilitated *folksonomies* or *collaborative tagging*, which are the classification and identification of content by the general public, rather than by domain experts (see Sections 2.4.3 and 3.6), and which facilitate serendipitous discovery of content [Vander Wal 2007]. Further, the tools and standards are readily available to combine content from multiple sources, even multimedia, in what are known as *mashups* (as in mashed up together). The cross-referencing inherent in geospatial data facilitate mashups and make them “one of the most powerful and ubiquitous bases for what is essentially a generalization of the concept of a relational join” [Goodchild 2008b].

Within *geographical information science* (GISc), user generated content is also known as *volunteered geographical information* (VGI) [Goodchild 2007b], see Section 4.5, and is made available as maps on public Web sites, such as Tracks4Africa [2016] and OpenStreetMap [2016], or as third-party data overlaid on *virtual globes*, such as Google Earth [Google 2016a]. Mobile electronic devices, particularly smartphones<sup>1</sup>, have increased dramatically the ability of people to generate and disseminate UGC.

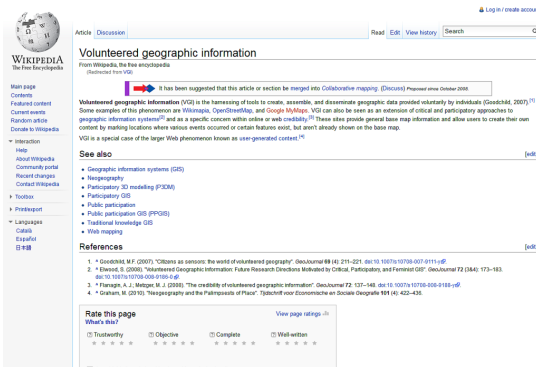
VGI is also contributed as observations to *citizen-science* projects (see Section 4.4), and as corrections to in-car navigation systems. Figure 1.2 shows examples of user-generated content on the Internet: *Wikipedia* with its entry on VGI, *Tracks4Africa* with its coverage of southern Africa, *OpenStreetMap* with its data of Cape Town (drawn using the Cycle Map renderer), and *Google Earth* showing part of the University of Pretoria with a user-contributed photograph from *Panoramio* [Panoramio 2016] of graffiti on the campus which, of course, is itself user-generated content.

A *virtual globe* provides masses of digital geospatial data over the Internet, typically in the form of a globe, and a *geobrowser* is the interface (or portal) to a virtual globe or other collection of geospatial data, typically allowing users to zoom into the data, switch data layers on and off, create three-dimensional views and add their own data, such as geospatial features (eg: roads and places of interest), tags (with text or links to Web sites) and photographs. Virtual globes are a major conduit for disseminating VGI, and hence

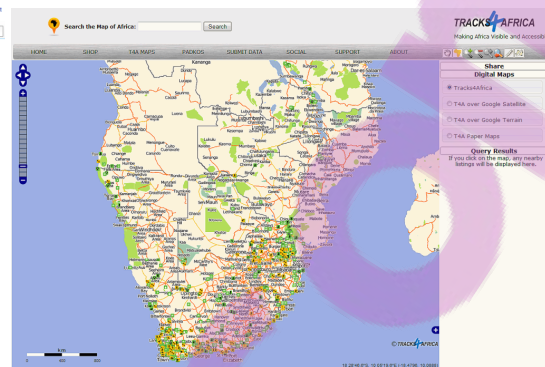
---

<sup>1</sup>A smartphone is “a mobile phone that performs many of the functions of a computer, typically having a touchscreen interface, Internet access, and an operating system capable of running downloaded apps” [Oxford 2016].

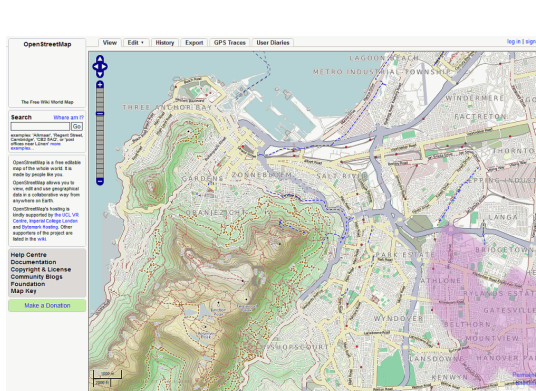
## 1. Overview of this thesis



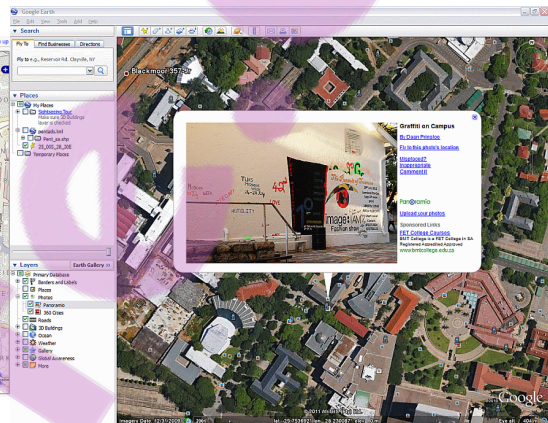
(a) Wikipedia



(b) Tracks4Africa



(c) OpenStreetMap



(d) Google Earth

Figure 1.1: Examples of online repositories of user-generated content

are closely coupled with VGI. They can also be perceived to be competition for SDIs. Virtual globes and geobrowsers are discussed in Section 2.10.

Some (but not all) user-generated content takes the form of contributions by the public-at-large to science, which is often called *citizen science* or *public participation in scientific research (PPSR)*. Much citizen science is routine or mundane, as discussed in Section 4.4, while some can be dramatic, such as the co-discovery of comets by the amateur astronomers David H Levy (particularly of Comet Shoemaker-Levy 9, which collided with Jupiter in 1994) and Thomas Bopp (particularly of Comet Hale-Bopp, perhaps the most widely observed comet of the 20th century) [Wikimedia 2016]. On the other hand, not all citizen science is user-generated content, as is discussed in Section 4.4.2.

A recent trend is to extend out-sourcing beyond traditional procurement, where the contractual relationship would invariably be initiated, if not actually completed, before any of the services are provided. Now, some organisations solicit completed services, rather than just offers to provide services, from the population at large, or the *crowd*: commonly known as *crowd sourcing* [Howe 2006], see Section 4.6. Crowd sourcing could produce user-generated content, but often does not (eg: soliciting from professionals), and user-

generated content is often not crowd-sourced, eg: blogs<sup>2</sup>.

Crowd sourcing, citizen science and UGC/VGI are sometimes confused with one another, see Section 4.2 and Table 4.1. For example, according to Li *et al* [2013], harvesting user-generated content is crowd sourcing. The differences are explained in Sections 4.2, 4.3, 4.4, 4.5 and 4.6. Chapter 7 reports on a survey of perceptions over some of these issues.

The term *neogeography* has been used for over 90 years to refer to new and emerging fields in geography [Haden 2008], which have obviously varied over the decades. Currently, the term seems to be applied primarily to the use of technologies such as GIS, Web mapping and navigation devices by anyone (that is, not just professional geographical information scientists — ie: VGI), and innovative, colloquial, *ad hoc*, unconventional, collaborative, integrative and/or open applications, mapping and/or data. Clearly, there are many different interpretations of the term *neogeography*, as one would expect. While some might infer that neogeography is “beyond” the professionals, much of the contribution is actually from professionals in the field, such as for critical GIS, qualitative applications and ethical issues. So, neogeography is not VGI and VGI is not neogeography, though there are obviously overlaps, as discussed in Section 4.7.

*Formal concept analysis (FCA)* is a formalism that can be used for classifying data or for examining classification systems, see Chapter 9. Essentially, FCA uses a lattice (a partially ordered set) of formal concepts with objects and attributes, and the linkages between them, for data analysis, knowledge representation and information management [Priss 2006]. The stability of a lattice indicates the extent to which objects and attributes can be removed without altering the lattice significantly: this is not a measure of the robustness of the lattice, but of the value or usefulness of the objects and attributes. I show how stability can be used to assess a taxonomy: specifically in Section 9.6, taxonomies of UGC and citizen science are assessed using various repositories of VGI.

Finally, in terms of providing the context for this research, are:

- **Inter-networking**, which is now beyond the control of any organisation or country and which has enabled social media services and social networking but also raises concerns over privacy, censorship, liability, the rights to exploit content, curation and the digital divide, see Chapter 3;
- **Metadata**, which is a structured and detailed description of a data element, data set, collection, Web service, process, product or other resource, to make the resource understandable and shareable into the future, see Chapter 5;
- **Quality**, whether inherent, measured and/or documented, which can apply to a data element, data set, collection, Web service, process, product or other resource, see Chapter 6;
- **Classification**, which is a logical but subjective grouping of things (possibly with ranking and/or in a hierarchy) according to their selected characteristics, to manage

---

<sup>2</sup>A blog (from ‘Web log’) is “a regularly updated Website or Web page, typically one run by an individual or small group, that is written in an informal or conversational style” [Oxford 2016].

## 1. Overview of this thesis

large amounts of data, describe the things and predict aspects, see Section 2.4 and Chapter 8; and

- **Standards**, which underlie metadata, quality, classification and other aspects of this thesis, see Section 3.13.

Inter-networking, metadata, quality, classification and standards are all part of the value chain from the content (including VGI) to an SDI, see Figure 1.3. The need for this exposition of the nature of VGI and its suitability for integration into SDIs is given below in Section 1.3.

## 1.3 Statement of the problem

### 1.3.1 What the problem is

A spatial data infrastructure is typically established by a government or public entity to facilitate managing and sharing geospatial data, primarily from official sources. However, an SDI is an evolving concept, particularly concerning the data, services, metadata, products and standards it consumes and provides.

Figure 1.2 illustrates that an SDI can be populated with geospatial content (data, services, products, etc) that is user-generated (ie: volunteered geographical information), from commercial sources (air survey companies, surveyors, mapping companies, GIS companies, data vendors, service providers, etc) and/or from official sources (national mapping agencies, national statistical agencies, other government departments, provincial and local government, etc).

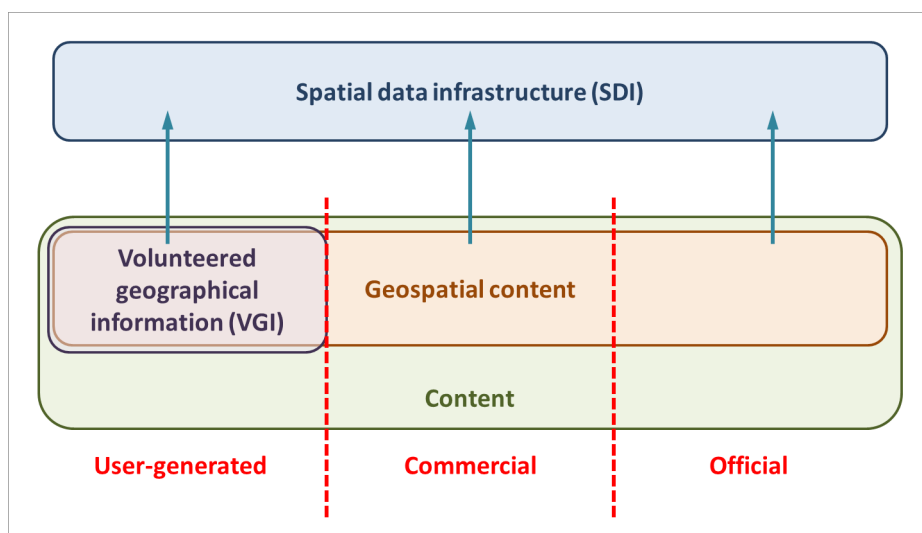


Figure 1.2: Geospatial content for an SDI can be official, commercial and/or VGI

The pervasiveness, power and low cost of inter-networking, social media, virtual communities, applications and mobile devices enable ordinary people (even the illiterate)

to document their environment, experiences, perspectives and prejudices, to share them widely and rapidly, and to query the content provided by anyone else — including official sources. Some user-generated content is of limited value, even to its creator (such as badly blurred and under-exposed photographs of some forgotten revelry).

However, much UGC can be very valuable and much VGI can contribute successfully to an SDI and as citizen science. VGI can extend the reach in time and space of official mapping agencies and the like, because of the sheer volume of humans and their devices, acting together or independently as sensors, recorders and disseminators. Further, VGI, repositories of VGI, virtual globes, mashups, folksonomies, crowd sourcing and neo-geography can challenge the traditional business models of official SDIs. Yet, metadata, quality, classification and standards can be challenges for VGI.

The problem is for VGI, separating the wheat from the chaff: identifying what is useful and what is spurious, which depends on the user's perspective and application (such as for an SDI) — a very difficult problem.

### 1.3.2 Why the problem is important

VGI can contribute to SDIs worldwide, at the local, national, regional and global levels. South Africa's National Development Plan (NDP) states in its chapter on transforming human settlement and the national space economy that:

*"There are pronounced limitations on citizen action at individual and community level. Although IDPs<sup>3</sup> are required to be participatory, engagement in planning processes and joint problem-solving often happens at a superficial level. Participatory processes are often formulaic and compliance-driven, and there are few incentives for citizens to engage in community-building. Citizen dependence on a state with limited capability leads to confrontational protests by individuals who are waiting for the state to provide houses and services. Simply providing the opportunity for local communities to take part in preparing their own plans may create new forms of inequality as better-resourced communities are far more likely to respond to the opportunity. A differentiated approach to spatial planning is required which allows simple approaches to be adopted in uncontested areas but provides for mechanisms to address the conflicts that are likely to emerge in other cases more speedily than at present" [South Africa 2012, pg 275].*

To address this, the NDP has as Objective 48, "Establish a national observatory for spatial data and analysis". Such an observatory "would collect, continually update and analyse data and other information relevant to spatial planning" [South Africa 2012]. A spatial observatory is part of an SDI, with its own value-add data sets and with data sets, analyses and other services that it can provide through the SDI as well as directly to its participants. The NDP's national observatory should form part of the South African Spatial Data Infrastructure (SASDI), which is discussed in Section 2.2.3.

---

<sup>3</sup>Integrated development plans.

## 1. Overview of this thesis

---

It can take a long time to establish an SDI and there is competition already, such as from virtual globes, open data repositories and other SDIs. Some SDIs have failed [Makanga & Smit 2008] or could be considered to be zombies, consuming resources without delivering much of value [Harvey *et al* 2015]. Hence, any SDI has to offer a valid value-proposition to justify its existence, including cooperating with other systems and organisations [Cooper 2013]. Depending on the circumstances, VGI can be as good as, if not better than, official data sets, see Haklay [2010] and Du Plooy [2012], for example. Whatever the source, though, the legibility and interpretability of the geospatial data (as maps, legends, driving instructions, etc) are increasing in importance because of the greater speeds at which they are used, and the greater volumes of data that need to be processed to provide clarity [Cochrane 2014].

Whilst current (up to date) remotely sensed data are generally available, this does not necessarily apply to other important geospatial data, such as addresses, house numbers, place names and points of interest. In developing countries where local governments tend to be less efficient, and there are informal settlements and rapidly changing human settlement patterns to make it even more difficult for them, it can be a massive challenge to have accurate and up-to-date geospatial data. However, ordinary citizens who live out their daily lives in these streets and villages can have a huge impact on the accuracy and availability of up-to-date spatial data by contributing VGI [S Coetzee 2011, *pers comm*].

Hence, as VGI can contribute to the likes of SASDI and the NDP's national observatory, because there are different types of VGI and to facilitate integrating VGI into SDIs, this research aims at understanding VGI in the context of the Internet, the World Wide Web and SDIs. For this, it investigates taxonomies of VGI and how well they apply to repositories of VGI; UGC in general; and geospatial data, quality, metadata, standards and perceptions of VGI. I also use formal concept analysis to assess how well several taxonomies of UGC and citizen science discriminated between various repositories of VGI.

Being new, all these concepts are not well understood: UGC, VGI, SDI, virtual globes, virtual communities, folksonomies, citizen science, crowd sourcing and neogeography. Even the more mature ideas of classification and metadata are not always implemented properly. As mentioned above, crowd sourcing, citizen science and UGC/VGI are sometimes confused with one another: the concepts can overlap, such as contributions to the Second South African Bird Atlas Project (SABAP2) [Animal Demography Unit 2016b], which are crowd-sourced, citizen science and VGI, see Sections 6.8.1 and 8.3.2.7. However, they can also be very different: tweets<sup>4</sup> are UGC, but neither science nor crowd-sourced, though they can be mined for scientific purposes. Contributions can be crowd-sourced from experts only (such as scientists) rather from the population at large, making the contributions neither citizen science nor UGC. Someone with an automated weather station can set it up to make its data available online<sup>5</sup>, which is VGI (if the location is known!) and citizen science, but not crowd sourcing.

Woldai [2002] suggests that an SDI initiative is doomed to fail without the “political will at the highest echelon of Governments fully committed” to involving citizens, investing

---

<sup>4</sup>A short message posted on the online social networking service, Twitter [Twitter 2016].

<sup>5</sup>Potentially illegal under the proposed South African Weather Service Amendment Bill [South Africa 2011]!

money, supplying logistics, investing in its people, establishing the relevant policies and legislation, and sharing data. Indeed, Makanga & Smit [2008] found that the two SDIs that existed in Africa in 2003 ceased functioning by 2008. Nevertheless, while based on his own experience, Woldai [2002]’s warning is speculative and there is much still to learn over what really makes an SDI successful. Hence, the survey of perceptions reported on in Chapter 7.

Ramsey [2006] observed that SDIs then were not succeeding well, particularly when compared to online commercial services such as Google Maps [Google 2016d] or Yahoo! Maps [Yahoo! 2016]. He suggested that the reasons were missing incentives for publishing one’s data, access speeds, reluctance to share data, the need for metadata (perpetuating the myth that metadata is only for other people and of no value to the data set’s creator) and that loosely-coupled systems<sup>6</sup> fail easily because they depend on systems and services out of one’s control [Ramsey 2006]. Recently, Tansley [2014a,b] followed up, pointing out that the commercial mapping servers are not SDIs, but commercial services with a narrow focus. He also considers that an SDI should be delivering value-added data rather than base data and that an enterprise’s own business processes can benefit from serving up its data through an SDI, even an imperfect one. Obviously, technology has also changed significantly since 2006.

### 1.3.3 What was done in researching the problem

This thesis appraises geospatial data, classification, SDIs, virtual globes, UGC, VGI, citizen science, crowd sourcing, neogeography, UGC validity, metadata, quality, standards, inter-networking, controlling the Internet, privacy, exploiting content, social media, curation, the digital divide, taxonomies of UGC and VGI, formal concept analysis (FCA), and related issues. This research aims at understanding VGI in the context of the Internet, the World Wide Web and SDIs, to facilitate integrating VGI into SDIs. For this, it investigates taxonomies of VGI and how well they apply to repositories of VGI; and UGC in general, geospatial data, quality, metadata, standards and perceptions of VGI. This thesis explores the impact of volunteered geographical information on spatial data infrastructures.

A *value chain* is “the process or activities by which a company adds value to an article, including production, marketing, and the provision of after-sales service” [Oxford 2016]. Figure 1.3 shows the high-level value chain for an SDI, from the content to the SDI. This figure shows that it is not just a matter of providing *content* to an SDI, but for the content to be of any value, it must be identified appropriately through a known *taxonomy* (ie: classified), must have its *quality* assessed, and must be documented (*metadata*). Further, *standards* are needed to support all aspects of the value chain from raw content to the SDI. *Inter-networking* has stimulated the creation of content and makes it available, as well as makes an SDI possible. However, inter-networking is not necessary for the classification, quality assessment and documentation phases, whether they are done manually, semi-automatically or fully automatically.

The research reported on here does not attempt to solve all of this problem of creating the

<sup>6</sup>An SDI is not necessarily loosely coupled.

## 1. Overview of this thesis

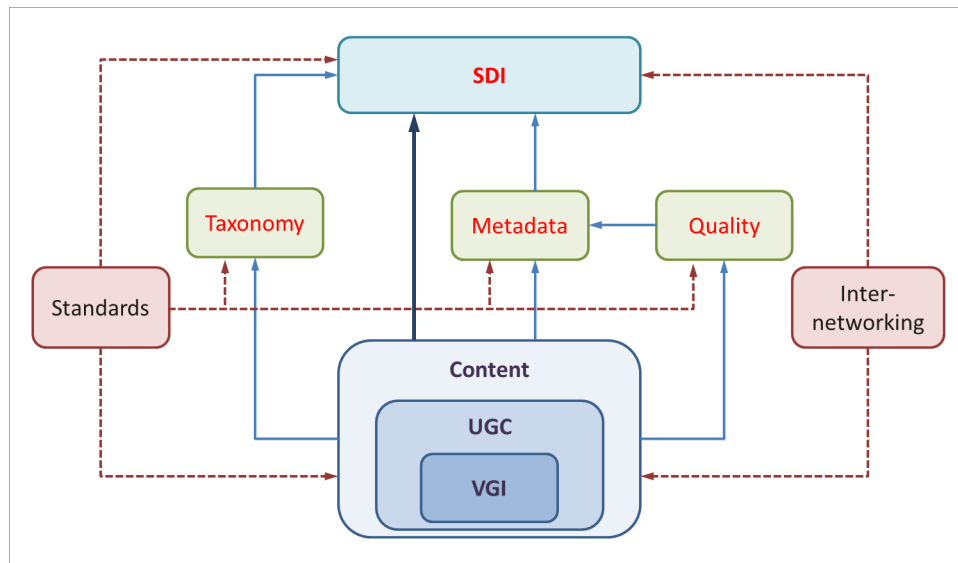


Figure 1.3: The value chain from content to SDI

SDI value chain, but the first six chapters provide the context and some building blocks for assessing VGI and repositories of VGI, such as for use in an SDI. These analyses can also be used for geospatial data in general, UGC or other content.

### 1.3.4 The contributions of this research

This thesis presents several original research contributions, summarised below.

- *To information science:*
  1. Section 3.8 explains why the Internet cannot be controlled, in spite of the best (or worst) intentions of those who should know better. Specifically, there are many alternatives for communicating, finding content, hiding content, power, funding, and so on.
  2. Sections 4.3 to 4.8 clarify the differences between user-generated content, volunteered geographical information, citizen science, crowd sourcing and neo-geography, because it appears that they are often confused with one another.
- *To geographical information science:*
  3. Sections 8.4 to 8.6 provide a qualitative assessment of several taxonomies of VGI against a selection of repositories of VGI. This showed that some of them could distinguish uniquely between the selected repositories, but not all. Some suggestions are then made for improving these taxonomies.
  4. Sections 9.4 and 9.6 show how formal concept analysis can be used for assessing existing taxonomies, such as to determine their discrimination adequacy.

Previously, FCA has been used to create taxonomies (eg: Kourie & Oosthuizen [1998]), but not to assess them.

5. Section 6.3 and Figure 6.1 describe the four stages for the recognition of the quality of a resource in general.
6. Chapter 7 reports on the results of a survey of geographical information professionals concerning their perceptions of VGI, SDIs and virtual globes, which was published in Cooper *et al* [2010a]. The questionnaire used is included in Appendix A.

- *To theoretical computer science*

7. In formal concept analysis, Sections 9.2.2 and 9.4 show that there can be value in instability in a lattice when assessing a taxonomy [Cooper *et al* 2010b], as the instability represents extreme values rather than noise.
8. Section 9.2.5 presents several lemmas for FCA on stability in a lattice. Amongst others, these provide lower and upper bounds for intensional and extensional stability indices.
9. Section 9.5 presents *stability exploration*, a technique for FCA, and a specification of it. Stability exploration can possibly be used as a decision support tool, see Section 9.5.4.

In addition, I also made several other contributions, summarised below. Some of these are original and some are synthesised from the existing literature.

1. Section 2.3.5 describes a comprehensive model of a feature.
2. Sections 2.4.5 and 2.4.6 provide details of the dangers of classification and of clever codes.
3. Section 3.9.1 provides an overview of privacy issues.
4. Section 3.9.3 highlights critical problems with the Protection of State Information Bill [South Africa 2013c] that do not appear in other analyses of the Bill. It draws on an invited presentation Cooper [2011a].
5. Section 3.12 exposes some myths on the causes of the digital divide.
6. Section 4.3 identifies the nature and various aspects of user-generated content.
7. Section 4.4.2 expands on the typology of citizen science of Wiggins & Crowston [2011].
8. Section 4.5.2 identifies various aspects of volunteered geographical information.
9. Section 4.6.3 goes beyond the commercially-oriented taxonomy of Saxton *et al* [2013] in identifying other types of crowd sourcing.
10. Section 4.8.2 identifies the types of blogs, and how these types correlate to volunteered geographical information.

## 1. Overview of this thesis

---

11. Section 4.8.3 identifies problems with assuming that traditional scientific media are inherently of a high quality.
12. Section 5.7 provides a comprehensive summary of the different categories or types of metadata.
13. Sections 6.5 and 6.6 present the dimensions and sub-dimensions of quality cohesively.
14. Sections 6.7.2 to 6.7.4 identify some challenges for VGI, which were included in Cooper *et al* [2011a].
15. Three VGI repositories are assessed against the quality dimensions and quality challenges in Section 6.8. A preliminary version of this analysis was published in Cooper *et al* [2012a].
16. Section 6.9.2 and Figure 6.9 show for a selection of repositories of VGI, the responsibility for their specifications mapped against the types of data they contain. This was also included in Cooper *et al* [2011a].
17. Section 8.3 identifies and assesses representative repositories that contain VGI of various types and to varying extents.
18. Sections 8.3.3 and 8.4 then provide quantitative assessments of these taxonomies, using these repositories to illustrate various aspects.
19. Section 8.7 presents an attempt at a taxonomy of user-generated content.
20. Section 9.3 shows a correlation of the feature model with formal concept analysis.
21. Section 9.6 shows an assessment using FCA of how well several taxonomies of user-generated content discriminated between various repositories of volunteered geographical information. A preliminary version was published in Cooper *et al* [2012b].
22. Sections 9.6.3 and 9.6.5, and Figures 9.15 and 9.19, describe how the ability to show sub-contexts in the FCA tool ConExp [Yevtushenko *et al* 2003], by selecting and deselecting attributes and objects, could be used to find manually more effective combinations of attributes, that is, the classes of the taxonomy being assessed.

## 1.4 Background

### 1.4.1 The literature reviewed and used

This thesis does not have a separate chapter on the literature survey, as it makes more sense to embed the references to the literature throughout the thesis, in the relevant places.

I have found Wikipedia [Wikimedia 2016] to have been a very valuable resource for getting to understand the concepts and issues discussed in this thesis, for formulating definitions or descriptions, and for locating references and other resources. I have also used

Wikipedia and such resources as the sources for a few paragraphs providing background information, particularly when they encapsulate well what is general knowledge. Traditionally, the use of user-generated content such as Wikipedia as sources is discouraged in academia, so I have been conscious of going beyond Wikipedia for my sources. Please see Section 4.8.1 for a discussion on using user-generated content. The result is that I have included few citations to Wikipedia — though this is true of other general references that have shaped my thinking and provided leads, such as other encyclopaedias and dictionaries. Okoli, Mehdi, Mesgari, Nielsen & Lanamäki [2012] have identified over 450 scholarly studies of Wikipedia. They conclude that Wikipedia has “content of considerable quality and quantity”. Indeed, Wikipedia has been so successful that it is contributing to the *digital divide* (see Section 3.12) by displacing up-to-date paper encyclopaedias, particularly in local libraries [Attwell 2013].

Nevertheless, I do appreciate the importance of the peer-reviewed literature and I have been careful not only to include references to peer-reviewed literature, but also to use them. The depth and breadth of the references is shown by at least 17 citations being to journals with impact factors over 15 (*Nature*, *Science*, *Behavioral and Brain Sciences* and *PLoS Medicine*), 12 to ACM and IEEE journals, 5 to the *South African Journal of Science*, and over 30 to the leading journals in GISc, such as the *ISPRS Journal of Photogrammetry and Remote Sensing*, the *International Journal of Applied Earth Observation and Geoinformation*, *Computers, Environment and Urban Systems*, *International Journal of Geographical Information Science*, *International Journal of Digital Earth*, *International Journal of Remote Sensing* and *Transactions in GIS*. Further, the citations range from 1738 to 2016.

The nature of scholarly publishing is also changing, with differing interpretations of what is meant by *peer review* (eg: making drafts available online for anyone to comment on) and what is *sufficient* peer review. Further, there are interesting, often online, peer-reviewed journals (eg: *D-Lib Magazine* [D-Lib 2016] and *First Monday* [First Monday 2016]) that do not appear in the indexes accredited by the likes of South Africa’s Department of Higher Education and Training, particularly Journal Citation Reports (JCR) and International Bibliography of Social Sciences (IBSS). This thesis deals with contemporary issues, such as surveillance and privacy, so includes many citations to news stories, particularly of respected sources such as the British Broadcasting Corporation (BBC), and The Guardian and Washington Post<sup>7</sup>.

#### 1.4.2 Geospatial terminology

To provide a background essential for understanding SDIs and VGI, their nature is analysed later in the next five chapters, together with the technologies, policies and concepts necessary for them. There is a plethora of terms used for *geospatial data*, such as “geospatial”, “spatial”, “geographical”, “geographic”, “geo-referenced”, “geographically referenced”, “geo-information”, “land” and “cartographic”. The differences between the terms were really significant in the 1980s and early 1990s, splitting the field into different groups that tended not to draw on the research and developments of each other [Cooper

---

<sup>7</sup>Joint recent winners of the Pulitzer prize for public service for their exposure of the surveillance activities of the National Security Agency [Pilkington 2014].

## 1. Overview of this thesis

---

1993]. However, these different terms are generally used interchangeably now, with the result that this thesis is about volunteered *geographical information* and *spatial data* infrastructures! For this thesis I shall generally use the term *geospatial*, as explained below in Section 2.3.2, where all these terms are discussed.

The literature on SDIs dates back to 1990 [National Academy of Sciences 1990] and much has been published on the topic, particularly since the Executive Order in 1994 establishing the national SDI in the United States of America [Clinton 1994]. This is discussed in more detail in Section 2.2 below.

The term VGI was introduced in 2007 by Goodchild [2007b] and already quite a bit has been published on it, especially in the context of an SDI (eg: Craglia *et al* [2008]; Budhathoki *et al* [2009]; Coleman *et al* [2009]; McDougall [2009]; Devillers *et al* [2012]; Coetzee *et al* [2013b]; Rak [2013]; Adams & Gahegan [2014]; Christensen *et al* [2014]; Cinnamon [2015]). This is discussed in more detail in Section 4.5 below. To give the context for VGI, this section is preceded by discussions on UGC in Section 4.3 and citizen science in Section 4.4, and followed by a discussion on crowd sourcing (which is often confused with UGC) in Section 4.6.

As discussed below in Section 4.5.1, there is concern over the use of labels such as *volunteered geographical information* or *VGI*. I felt that it would be useful to find out what perceptions were held about VGI and related concepts, and Chapter 7 discusses the results of a questionnaire on such perceptions. This dichotomy over the meaning of VGI is explored further in Chapter 8, where published taxonomies of user generated content are assessed qualitatively, and in Chapter 9, where they are analysed using formal concept analysis [Wille 1982]. VGI is discussed in more detail in Section 4.5 below.

*Metadata* is often defined narrowly as *data about data* (eg: [ISO 19115 2003]), but metadata is more than just that. Metadata also describes processes, services, systems, etc — their *provenance* — which is why the new definition of **metadata** for the revised standard is *data describing resources*, with a **resource** being an *identifiable asset or means that fulfils a requirement* [ISO 19115-1 2014]. Metadata, whether declared or inferred, is crucial for understanding the value or usability of data, whether from official sources or from users. Unfortunately, many make assumptions about the data (ie: they infer the metadata and quality) that can be false: the well-known problem of *Garbage In, Gospel Out* — *GIGO*! These issues are discussed in Chapters 5 and 6. With the exposure of the widespread surveillance by the United States of America, the term *metadata* is now much more widely known — and has taken on a sinister meaning [Wise & Landay 2013]!

### 1.4.3 Limitations with the Internet

The first phase of the World Wide Web, *Web 1.0*, had network resources, information and services delivered and developed only by programmers and the administrators of Web sites. Users were only passive receivers of what was delivered over the Web and had no major impact on the content. The next phase, *Web 2.0*, largely revolutionized the perception of the Internet, seeing the rise of social networking portals and mechanisms for publishing content on the Internet without specialized knowledge. Now, anyone could

---

## 1. Overview of this thesis

---

become a provider of information on the Web. Web 2.0 is often characterized as a transition from a *read-only Web* to a *read-write Web*, with users no longer just passive consumers, but also creators of resources [Lessig 2005; O'Reilly 2005; O'Reilly & Battelle 2009; Cooper *et al* 2011a; Swartz 2013]. “Web 2.0 has a post-modern feel with its emphasis on engaging the individual, on personalization of content, and on the subjective side by side with the objective”; further, “definitions of the Geospatial Web or GeoWeb typically emphasize the power of geographic location as a key for integrating knowledge, and for providing context” [Goodchild 2008b].

Unfortunately, one of the consequences of the Internet and the World Wide Web (WWW) is the massive explosion in raw data available to anyone with a computer connected to the Internet — far too much for any human to manage or absorb. This has led to the development of portals and search engines to help users find relevant information. However, with them also comes the *filter bubble* [Pariser 2012], whereby based on one's previous activities on the Web, algorithms decide what one is exposed to and in what order: while such *personalization* generally enhances one's use of the Web, it “is showing us what it thinks we want to see, but not necessarily what we need to see” [Pariser 2012]. One example of this is the update to Google Maps [Google 2016d] in 2013, which personalizes maps based on user queries and other data Google has on the user: guiding rather than enabling exploration [Ball 2013]. This might well violate Ranganathan's Third Law: “every book its reader”, concerning context rather than just raw content [Ranganathan 1931]. That is, many Internet users do not know their real requirements or what resources are available, so the job of librarians (and hence search engines and other Web applications) is to enable browsing, linking and hence serendipity, by helping the “resources find the people who want and need them the most” [Johnson *et al* 2009].

A further problem with the filter bubble is that if one is ignorant of key aspects, such as the history of the topic or the required scientific understanding, one could unwittingly still be confined to the filter bubble, even if one thinks that one is exploring widely. For example, Simkin [2014] refers to the then Wikipedia article on the Dewey Commission into the show trials of Trotskyists in Moscow in 1936/7. While Simkin [2014] does not consider the article to be actually lying, he considers that the article takes a pro-Stalinist stance, omitting much anti-Stalinist material — such as the KGB archives that showed that the Commission's report was accurate, that there was not a Trotskyist-Nazi-Western plot, that the condemned were innocent, that those who did not ‘confess’ in open court were simply shot without trial, and that after the Great Purge, Stalin ordered the murders of all the NKVD officers who organised and conducted the executions [Simkin 2014]. It is perhaps easy to get paranoid within a bubble: the bubble also seems to promote polarization, that is, moving the group to more extreme viewpoints [Hendricks 2014].

Unsurprisingly, the filter bubble effect did not start with the Internet and was probably much worse centuries ago, when most people travelled little and the elite controlled what little media there was. For example, the Commonwealth's licensed newsbooks in England in 1649 “could suppress or modify information just as much as they could spread it” [Poyntz 2009].

For an example of a limitation with searching (which might have been caused by a filter bubble, or some other cause), see Section 4.9.1. The result is that one can get an increas-

## 1. Overview of this thesis

---

ingly narrow perspective of the world (in spite of getting more and more information), or *confirmation bias*, which I would suggest will reinforce differences and prejudices. Such selective exposure can be deliberate, of course, and in their study, Cloonan & Dove [2005] found that this is even done by the highly educated. “*Technology celebrates connectedness, but encourages retreat ... The more distracted we become, and the more emphasis we place on speed at the expense of depth, the less likely and able we are to care*” [Foer 2013].

The effects of this can be seen in the auto-completion offerings of search engines and other services when one starts typing in a text input box. These are limited by their complexity and required response times, so the proffered options tend to reflect the common queries made, which can reproduce stereotypes and prejudice [Baker & Potts 2013]. This reinforces the filter bubble, though Mazières & Huron [2013] suggest that it could also “induce serendipity through surprising or complementary propositions”. On the other hand, the suggestions made by an auto-completion service can also be used for analysing cultural trends [Mazières & Huron 2013].

The mass of raw data is also the motivation behind the development of the concept of the *Semantic Web* [Berners-Lee *et al* 2001; Bizer *et al* 2009], see Section 3.5. Indeed, with all the undifferentiated and unverified data available, I would suggest that this is not the information revolution — it is the *nonsense* revolution! These issues are discussed in more detail in Chapters 3 and 4 below.

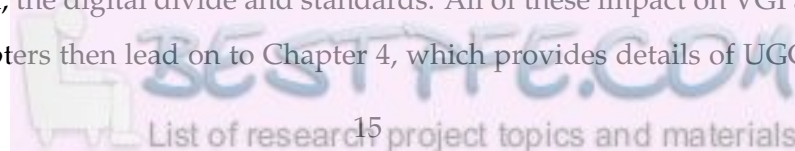
## 1.5 Summary and looking ahead

This chapter has introduced the research described in this thesis and given a brief overview. Figure 1.4 shows how the chapters in this thesis link together to provide the “story line” for this research. For reference, this picture is repeated at the top of the first page of each chapter, with the relevant chapter highlighted. The following chapters provide the exposition of the nature of VGI and its suitability for integration into an SDI.

Chapter 2 provides details of SDIs, formal models of SDIs, and the terminology, types, complexities and models of geospatial data, which leads on to Chapters 4 and 7. It introduces quality and metadata (detailed in Chapters 5 and 6), discusses classification and also touches on incremental updating and versioning, cartography, virtual globes and geobrowsers. This chapter provides the context for understanding SDIs, VGI and how VGI can contribute to an SDI; and the context for the assessment of repositories of VGI, done in Chapters 8 and 9.

Chapter 3 provides details of the context that made the proliferation of user-generated content and volunteered geographical information and the development of spatial data infrastructures possible, and the impact of such fecundity: inter-networking (which is much more than just the Internet and the World Wide Web), online services and content, the Semantic Web, social media, social mapping, (impossibility of) controlling the Internet, open archives and access, privacy, censorship, liability, patents, copyright, open access, curation, the digital divide and standards. All of these impact on VGI and SDIs.

These two chapters then lead on to Chapter 4, which provides details of UGC and VGI,



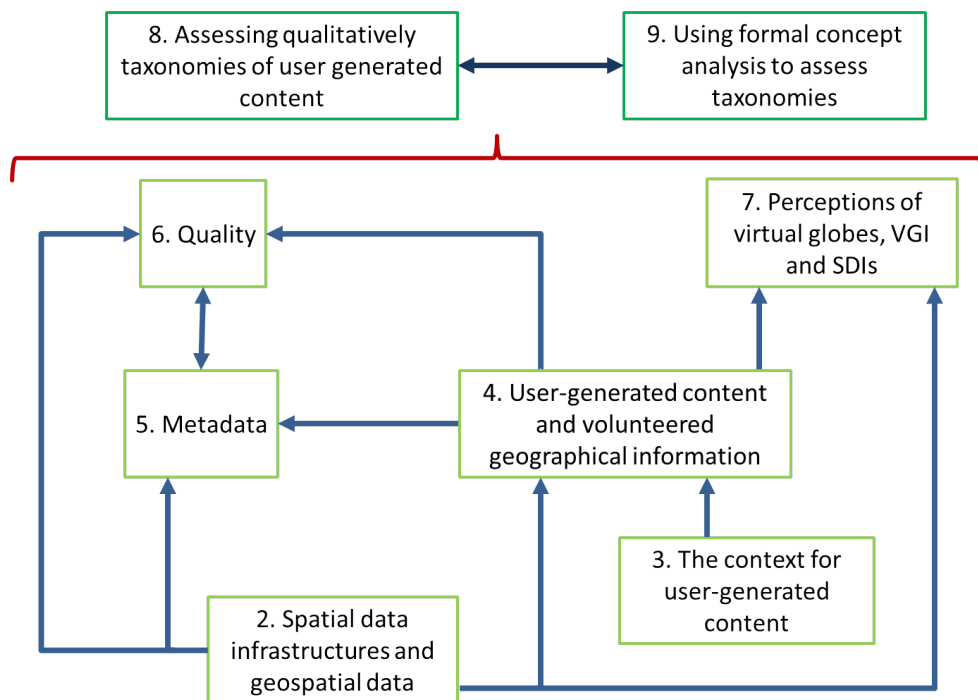


Figure 1.4: Overview of how the chapters link together

for which the quality and metadata are discussed in Chapters 5 and 6. But to understand UGC and VGI, Chapter 4 also provides details of concepts with which they are often confused, namely citizen science, crowd sourcing and neogeography. It also discusses the validity of using UGC in scholarly research, the quality of the traditional scholarly media, how traditional scholarly media matches official producers of geospatial data, and citing UGC, data and data repositories.

Chapters 5, on metadata, and 6, on quality, are closely interlinked, and both are key characteristics of VGI repositories, which are assessed in Chapters 8 and 9. Further, the most common objections raised against VGI are perhaps the uncertainty over the quality of the VGI and of the documentation of the data, that is, the metadata.

Specifically, Chapter 5 covers the definition, aspects, encoding, tools, categories, standards and limitations of metadata, as well as how a specification is the inverse of metadata. It also considers metadata against searching, as linked open data, and for VGI. Chapter 6 discusses in detail the aspects, recognition stages, dimensions of quality, and standards for geospatial data. It illustrates the quality recognition stages using typical problems with global navigation satellite systems (GNSS). Chapter 6 also assesses three VGI repositories against the quality dimensions, classifies types of VGI from the perspective of quality, and considers quality challenges for VGI.

Chapter 7 explores the current understanding of VGI, SDIs and virtual globes, through a questionnaire that I developed and used at meetings in Addis Ababa and Kempton Park. The questionnaire itself is included in Appendix A. This chapter, with the appendix, is

## 1. Overview of this thesis

---

essentially the equivalent of the journal article, Cooper *et al* [2010a].

These first seven chapters provide the context for the next two chapters, which are closely interlinked. Chapter 8 presents an overview of taxonomies of UGC and citizen science, and of various repositories of VGI and their characteristics. It then does a qualitative assessment of the repositories and of the taxonomies to discriminate adequately between the repositories. The three repositories assessed against the quality dimensions in Chapter 6 are also assessed here and in Chapter 9. Finally, this chapter presents my preliminary attempt at a taxonomy of UGC.

Chapter 9 introduces *formal concept analysis (FCA)*, including stability and instability in a lattice, tools that support FCA and attribute exploration. It includes original contributions that I have made in discovering *stability exploration* and on the value of instability in a lattice, such as the rationale for stability exploration, a methodology for implementing it, some possible applications of stability exploration, some lemmas on stability in a lattice, and the correspondence between FCA and the feature model (see Chapter 2). Finally, FCA is then used in the chapter to conduct a more rigorous analysis of the published taxonomies described and used in Chapter 8, against the repositories also described there, covering discrimination adequacy, absent and redundant attributes and objects, and high intensional and extensional stability. Please note that this chapter will be difficult to read for those without a background in FCA, but it is a critical part of my thesis.

Chapter 10 concludes this thesis, describing enhancements that can be made to cater for VGI and describing future research topics. This includes those questions for further research that are posed in some of the other chapters.

Appendix A provides a copy of the questionnaire on VGI and virtual globes, the results of which are discussed in Chapter 7. For ease of reference, Appendix B provides the details of the five taxonomies of UGC discussed in Chapter 8.

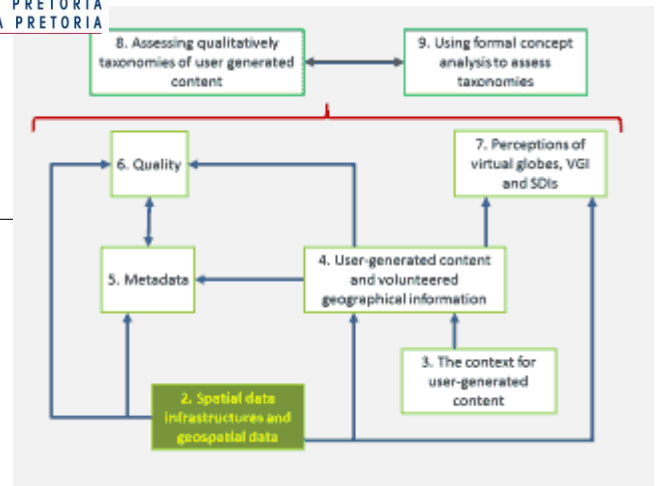
This thesis concludes with the bibliography of references cited in this thesis and the colophon. Traditionally, a colophon has been “a tailpiece in a manuscript or book, often ornamental, giving the writer’s or printer’s name, the date, etc” [Oxford English Dictionary Department 1973], but also providing details of fonts used, the production process (such as  $\text{\LaTeX}$  style files and software tools used), versioning, how cross-referencing was done, and so on. I also use the colophon to explain my particular usage of English. While it might appear idiosyncratic to some, I think that my grasp of English is far better than that of most. Hence, the reader should refer there should they have concerns over my style!

The appendices form an integral part of this thesis. They have been placed at the end of this thesis for ease of reference.

\*\*\*\*

---

*1. Overview of this thesis*



## Chapter 2

# Spatial data infrastructures and geospatial data

### 2.1 Overview of the chapter

This chapter discusses spatial data infrastructures (SDIs), geospatial data and models, classification, cartography and related concepts. Specifically, this chapter covers the following.

- Section 2.2 discusses *spatial data infrastructures* in South Africa and elsewhere, to explain the SDI concepts and terminology, providing the context for understanding how volunteered geographical information can contribute to an SDI, and the progress with SDIs in South Africa.
- Section 2.3 discusses the terminology, types and complexities of *geospatial data* or geographical information, including a comprehensive model of a **feature**.
- Section 2.4 discusses *classification* in its different forms, folksonomies, ontologies, problems with classification and encoding classifications, including the curse of clever codes, because these are used for assessing the repositories of VGI in Chapters 8 and 9.
- Section 2.5 provides an overview of *models for geospatial data* in a geographical information system (GIS). These are needed to understand VGI and SDIs, due to the diverse terminology and concepts used, which can be confusing.

---

## 2. Spatial data infrastructures and geospatial data

- Section 2.6 discusses *formal models of SDIs*, particularly the work on the Commission on Geoinformation Infrastructures and Standards of the International Cartographic Association (ICA). Such models provide useful frameworks for understanding SDIs.
- Section 2.7 introduces briefly the concepts of *data quality* and *metadata*, which are discussed in detail in Chapters 6 and 5, respectively. The two concepts are closely coupled and are key characteristics of the repositories of VGI assessed in Chapters 8 and 9. Further, the most common objections raised against VGI are perhaps the uncertainty over the quality of the VGI and of the documentation of the data (eg: Cooper *et al* [2010a]).
- Section 2.8 introduces *incremental updating and versioning*, which is a key problem with maintaining SDIs and repositories.
- Section 2.9 discusses *cartography*, which is effectively the user interface to an SDI.
- Section 2.10 introduces *virtual globes and geobrowsers*, which can be seen as forms of an SDI, competition for an SDI, and/or sources of VGI.

While this chapter is mainly for setting the scene and hence does not make any major original contributions, the key contributions that I have made in this research that are presented in this chapter are:

- A comprehensive model of a **feature**, see Section 2.3.5; and
- Details of the dangers of classification and of **clever codes**, see Sections 2.4.5 and 2.4.6.

Further, this chapter raises some questions for further research:

1. In Section 2.2.2, to what extent will citizens be prepared to adhere to policies and other standards, on which they invariably have had no input? How can citizens be included in the development of policies and standards, whether or not they are yet VGI contributors?
2. In Section 2.4.5, can an expert hierarchical taxonomy provide more certainty than raw searching, or should one just dispense with taxonomies?

## 2.2 Spatial data infrastructure

### 2.2.1 The nature of a spatial data infrastructure

No national mapping agency (NMA) captures and processes by itself, all the geospatial data for its products. The NMA will obtain some data sets from the various custodians for those data sets. Generally, the NMA will also contract professionals to provide geospatial data. The NMA will need workflows and protocols for their various products for each of their data sources, including their in-house data capture and processing resources. The different types and sources of VGI should be able to fit into these workflows, perhaps

## 2. Spatial data infrastructures and geospatial data

at different stages. Unsurprisingly, such workflows and inter-institutional arrangements have evolved into broader collaborations, particularly as *spatial data infrastructures (SDIs)*.

An SDI is more than just the technology of a geographical information system (GIS). The term SDI generally refers to the collection of technologies, policies, standards and institutional arrangements that facilitate the availability of, and access to, geospatial data. It provides a basis for geospatial data discovery, evaluation and application for a variety of users and providers [Nebert 2004]. An SDI is an evolving concept about facilitating and coordinating the exchange and sharing of geospatial data and services between stakeholders from different levels in the geospatial data community [Hjelmager *et al* 2008]. Figure 2.1 is from Hjelmager *et al* [2008] and shows a high-level view of an SDI, modelled using the *unified modelling language (UML)* [ISO/IEC 19501 2005] *class diagram* through the *Enterprise Viewpoint* of the *Reference Model for Open Distributed Processing (RM ODP)* [ISO/IEC 10746-1 1998]; see also Section 2.6.

In Figure 2.1, an SDI is in the centre and is an *aggregation* (shown by the white or open diamond arrow heads) of one or more of each of policies, products, metadata, processing tools and connectivity. In return, the policies only exist with an SDI, but the products, metadata, processing tools and connectivity can exist without an SDI. The SDI's attributes are its scope and implementation plan. The connectivity has bandwidth as an attribute, uses technology and is used by processing tools. Each product has metadata and each set of metadata describes one and only one product. Each product can also consist of other products. The processing tools use the products and metadata. The box to the top right with dotted line connections is a *note* which states that the policies should support interoperability through the technology and processing tools.

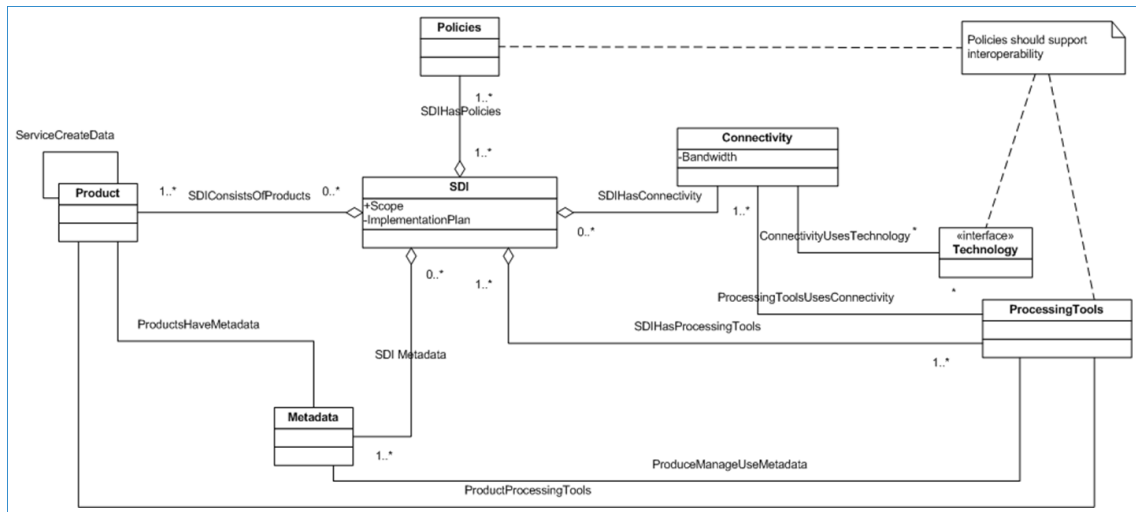


Figure 2.1: The high-level UML classes of the enterprise viewpoint of an SDI [Hjelmager *et al* 2008].

Hence, the SDI needs to include other technologies, such as archiving, connectivity and online services, and adhere to standards and protocols. The SDI also needs to cater for 'soft' issues, such as business models, cooperative agreements, legislation, marketing, ed-

---

## 2. Spatial data infrastructures and geospatial data

---

ucation and structures (such as committees) for coordination and management. One SDI can be part of another SDI, either functionally (such as a national water SDI within a general national SDI) or hierarchically (such as the Europe-wide SDI, INSPIRE (Infrastructure for Spatial Information in the European Community) [European Parliament 2007], which is based on the national SDIs of Member States [Cooper *et al* 2011c]. Kaczmarek *et al* [2014] feel that INSPIRE can really help spatial planning in Europe — provided that the individual countries contribute enough resources. Cooper *et al* [2014] report on a review we did at the CSIR on the SDIs in Australia, Brazil, China, India and South Africa.

In collecting various definitions of an SDI, Dessers [2012] noted that the definitions either do not mention the components of an SDI (such as technology or human resources), provide a general typification of the components, or provide a list of the components. Then, the definitions either do not mention any objectives of the SDI (such as data access), provide only data-related objectives, or provide both data-related and use-related (broader) objectives.

### 2.2.2 Volunteered geographical information contributions to spatial data infrastructures

Many countries are developing SDIs to manage and use their spatial data assets better by taking a perspective that starts at a local level and proceeds up through state, national and regional levels to the global level. This has resulted in the development of different forms of SDI at, and between, these levels.

Typically, an SDI is populated with data from government entities that have a formal mandate to provide, update and maintain spatial data (ie: data custodians) and that are required to adhere to government policies and legislation (such as South Africa's SDI Act [South Africa 2003], or the European Union's INSPIRE Directive [European Parliament 2007]), but that are also funded and mandated to fulfil these roles. These entities include not only the national mapping agencies (NMAs) and geodetic and cadastral surveying agencies, but also national and provincial government departments providing spatial data specific to their domain (eg: socio-economic statistics, water, health, environment or education), local authorities and other agencies (eg: the Earth Observation Centre<sup>1</sup> of the South African National Space Agency (SANSa), which is the main receiver, archiver and distributor of satellite imagery in South Africa).

Both the bottom-up approach to an SDI (also known as an inverse infrastructure), which involves all levels of government and the private sector, is user-driven and self-organising [Coetzee & Wolff-Piggott 2015], and the trend towards VGI increase the number of stakeholders in the SDI, their associated diversity and heterogeneity, and the resources at their disposal. They also raise the questions of accuracy and trust. While mandated organisations *should* produce data of higher quality and in greater bulk, their mandates and priorities (eg: the need to provide national coverage or the need to support a specific national priority) might result in significant delays before they update data in certain areas. On the other hand, the public at large might be the best available source to maintain the

---

<sup>1</sup>Formerly the CSIR's Satellite Applications Centre (SAC).

## 2. Spatial data infrastructures and geospatial data

currency (ie: keeping the data up to date) of local data, such as verifying street names and addresses, or documenting changes in the local spatial data — the question is whether or not they can be trusted to provide accurate data and to document what they have provided (ie: provide *metadata*). Also, to what extent will citizens be prepared to adhere to policies and other standards, on which they invariably have had no input? These issues are discussed below in more detail in Sections 4.3 and 4.5, which discuss *user-generated content* and *VGI*, and in Sections 6.2 and 5.2, which discuss *quality* and *metadata* respectively. Bravo *et al* [2015] found that the metadata elements provided by Wikimapia and OpenStreetMap matched most of those required by *Perfil MGB*, the metadata profile for the Brazilian SDI.

Conceptually, an SDI can exist without users, but VGI needs users, by definition! It is possible for an SDI to fail, such as by restricting the use of data (eg: for security reasons), ignoring the requirements of end users (as opposed to just those of the institutions tasked to provide data), having a faulty business model (eg: without adequate funding sources), lack of resources (funding, skills, equipment, connectivity, data, metadata, services, etc), or lack of cooperation from key stakeholders. For example, in 2003 there were two African countries with SDI clearing houses and in 2008 there were three — but the two from 2003 were no longer operational in 2008 [Makanga & Smit 2008]. An SDI can also become a zombie, consuming resources without really delivering anything of value: unread reports, duplicated spending, scope creep, unused metadata, etc [Harvey *et al* 2015]. Using VGI in an SDI highlights the importance of the user as a stakeholder, particularly for improving the SDI. An effective SDI should generate participatory VGI because it provides value to end users and hence stimulates them to contribute to the SDI [Cooper *et al* 2011c].

An official SDI will generally have a rigid, well-defined framework, whereas an SDI dominated by VGI could be fluid and unconstrained. VGI can be integrated into formal models of an SDI [Cooper *et al* 2011c]. The strengths of VGI include openness, market-orientation and interaction between stakeholders, while the weaknesses of VGI include heterogeneous data (eg: VGI coverage mainly where young and well-educated people live — creating a *digital divide* within countries, see Sections 4.3 and 3.12), lack of metadata (some contributors are anonymous) and uncertainty over the reliability of the data in comparison to official data, see Section 4.5. VGI contributions can also be “more intense where there are big transformations such as infrastructure projects” for the Olympic Games in 2016 [Borba *et al* 2015]. SDIs are evolving from a rigid traditional framework (of which there might be few left now) towards a mixed VGI model [Cooper *et al* 2011c]. One area where VGI can contribute to an official SDI now is through change detection for follow-up by professional staff, as is being explored by swisstopo [Guélat 2009] and South Africa’s NGI [Siebritz 2014]. While national mapping agencies have broader mandates than just collecting data and producing maps (eg: establishing and maintaining national reference systems, or authoritative control over private data), they do need to exploit the opportunities offered by VGI [Devillers *et al* 2012]. Land Information New Zealand (LINZ) has initiated a project in the form of a competition, to crowd-source from school children the capturing of building footprints from aerial photography. The University of Canterbury will be analysing the quality of the submitted VGI [LINZ 2016].

---

## 2. Spatial data infrastructures and geospatial data

### 2.2.3 Towards spatial data infrastructures in South Africa

Currently, the South African Spatial Data Infrastructure (SASDI) does not yet exist, though it is being developed and components are in place, as discussed below. Some provincial governments and local authorities in South Africa are also developing SDIs (though they might not term them as such) which would likely form part of SASDI, for example, the collaboration between the Provincial Government of the Western Cape and the City of Cape Town Municipality [Heald 2011], and Spisys, the spatial planning and information system of the Department of Rural Development and Land Reform for the Free State and Northern Cape, run in partnership with the two provincial governments [Spisys 2016].

#### 2.2.3.1 The Promotion of Access to Information Act

The **Promotion of Access to Information Act (PAIA)** [South Africa 2000] made data (including geospatial data) from South African government Departments readily available to all at nominal cost. Initially, the Departments were overwhelmed by the demand for geospatial data, but they have now established the required capacity and mechanisms for satisfying the demand. Fortunately, unlike some other countries (see below), this Act has not yet been watered down to impede the access to geospatial data. However, the Protection of State Information Bill [South Africa 2013c] is now of great concern in this regard, particularly as it allows *categories of information* to be classified, which means that content can be classified retrospectively [Cooper 2011a]. This Bill is discussed below in Section 3.9. PAIA has now been taken further for geospatial data by the Spatial Data Infrastructure Act [South Africa 2003], which should improve the availability of authoritative geospatial data even more, by facilitating the development of SASDI.

In the United Kingdom, for example, the number requests for information under their Freedom of Information Act that were refused has risen from 18% in 2005 (when the Act came into force) to 22% by the last quarter of 2009 [APPSI 2010]. In the United States, the US Justice Department recently withdrew their proposal to change their Freedom of Information Act which would have allowed the US government to claim that requested records do not exist, even if they do [Kravets 2011]. Further, the terms of the proposed Stop Online Piracy Act (SOPA) and Protect IP Act (PIPA) in the US were so draconian, that they triggered wide-ranging protests on the Internet on 18 January 2012, including the unprecedented blacking out of the English Wikipedia [Wikipedia 2012].

#### 2.2.3.2 The Spatial Data Infrastructure Act

The **Spatial Data Infrastructure Act (SDI Act)** [South Africa 2003] was signed into law at the beginning of 2004, but only starting to come into effect in mid-2010, for various reasons. The Act places requirements on each *data custodian*, which is an organ of state or a contractor “which captures, maintains, manages, integrates, distributes or uses spatial information” [South Africa 2003]. Clearly, this is quite a broad definition of *data custodian*, which will hopefully be refined to include only those responsible for the base (or fundamental) spatial data sets, once the custodians have been appointed officially. With

## 2. Spatial data infrastructures and geospatial data

---

funding from the then Chief Directorate: Surveys and Mapping (CDSM) of the South African Department of Land Affairs<sup>2</sup>, a project has already identified the fundamental data sets for Africa [Gyamfi-Aidoo *et al* 2006], and South Africa's do not differ dramatically from these. With funding from the Development Bank of Southern Africa, a project identified the base data sets for South Africa, with the project being executed by some who were involved in the African study [Schwabe & Govender 2012]. See Section 2.3.3 for a discussion on the types of geospatial data, including base or fundamental data.

### 2.2.3.3 The Committee for Spatial Information (CSI)

The **Committee for Spatial Information (CSI)** has been established in terms of the SDI Act, to implement the Act. I was appointed to represent the CSIR on the committee on 10 March 2009, but the notice appointing the members was gazetted only on 21 May 2010 in the Government Gazette [Nkwinti 2010] and the CSI met for the first time on 21 June 2010. The Directorate: National Spatial Information Framework (NSIF) in the Department of Rural Development and Land Reform is tasked with providing the secretarial and administrative support for the CSI. Unfortunately, though, as is obvious from the Department's name, geospatial data is not the core business of the Department. Further, in general the CSI members lack expertise on SDIs — unsurprisingly, as there is a general lack of knowledge of SDIs in South Africa. To address this, CSI and the Centre for Geoinformation Science at the University of Pretoria arranged a successful Spatial Data Infrastructure (SDI) Workshop on 30 May 2011 [CGIS 2011]. At its meeting on 15 March 2011, the CSI adopted its reference document and established six sub-committees: Policy and Legislation; Data; Systems; Standards (which I chair); Education and Training; and Marketing and Communication.

Unfortunately, even after being extended, the term of the CSI ended on 30 June 2014 and the new CSI was only appointed in March 2016 (though without key stakeholders) and first met on 6 July 2016. Nevertheless, whether unofficially or through other forums, work continued in any case to achieve the aims of SASDI, because it is to the benefit of the line business of enough government entities on all three tiers of government. To date, the CSI's achievements include the following, done primarily through its sub-committees:

- Completed a study to identify base data sets and their corresponding custodians, which was funded by the Development Bank of Southern Africa;
- Finalised policies on custodianship and pricing, that are awaiting approval by the Minister;
- Implemented a pilot for capturing and publishing of metadata, hosted by the South African Environmental Observation Network (SAEON);
- Initiated a data collection project register, to reduce duplication and improve efficiency and effectiveness of geospatial data collection;

---

<sup>2</sup>Now the Chief Directorate: National Geospatial Information (NGI) of the Department of Rural Development and Land Reform.

---

## 2. Spatial data infrastructures and geospatial data

---

- With the South African Bureau of Standards, established a site license for the relevant standards for the data custodians identified by the CSI, see Section 2.2.3.5;
- In collaboration with the Geo-Information Society of South Africa (GISSA), investigated the supply and demand of GISc skills nationally to address socio economic ills of the country, see Section 2.2.3.7; and
- Held workshops and other events for training on SDIs and for sharing experiences and discussing issues related to SDIs.

The CSI has also spawned the development of the *South African Geo-spatial Information Management Strategy (SAGIMS)*, because according to the SDI Act, the objectives of the SASDI include “promote effective management and maintenance of spatial information” and “create an environment which facilitates co-ordination and co-operation among all stakeholders regarding access to spatial information” [South Africa 2003]. SAGIMS is not just for the CSI and SASDI, or even for government, but is for the whole of the country [CSI 2014].

In fact, before the SDI Act was written, there should first have been a Green Paper<sup>3</sup> then a White Paper<sup>4</sup> on geospatial information management, to establish the policy framework for the SDI Act. The SAGIMS document might well become a Green Paper. SAGIMS needs to align with the country’s long-term developmental agenda, including the National Development Plan [South Africa 2012], which SAGIMS will support [Siebritz & Fourie 2015]. Further, SAGIMS needs to illustrate how the NDP’s national observatory should be built on top of SASDI (eg: see Coetzee & Smit [2015]), and how SASDI can support the successful implementation of legislation such as the Spatial Planning and Land Use Management Act (SPLUMA) [South Africa 2013d]. The work on SAGIMS is being done in four Commissions:

- **Data**, with objectives on availability and accessibility, reliability, relevance and usability;
- **Capacity and Capability**, with objectives on skills supply, skills demand and understanding the value of geospatial information;
- **Geo-Information Communication and Technology**, with objectives on a high-capacity network infrastructure, high-capacity computer storage and secure ICT environments; and
- **Policy and Legislation.**

So, even without an active CSI, much relevant work is being done in developing SAGIMS. Further, there are other initiatives such as *Mapping Africa for Africa (MAfA)*, an initiative of the United Nations Economic Commission for Africa (UN ECA) and the International Cartographic Association (ICA), launched by the *Durban Statement on Mapping Africa for Africa* on 16 August 2003 at the 12th General Assembly of the ICA [United Nations Economic Commission for Africa 2005]. One MAfA initiative was the project to identify the

---

<sup>3</sup>A discussion document or tentative report.

<sup>4</sup>An authoritative report presenting the preferences of government.

## 2. Spatial data infrastructures and geospatial data

---

fundamental geospatial data sets for SDIs in Africa and to conduct an inventory of them for each country in Africa [Gyamfi-Aidoo *et al* 2006; Schwabe & Govender 2010a,b].

A second MAfA initiative is the book, *Guidelines of Best Practice for the Acquisition, Storage, Maintenance and Dissemination of Fundamental Geo-Spatial Datasets: Mapping Africa for Africa (MAfA)* [Clarke 2014], which will be made available online for free, to encourage its use and the development of SDIs across Africa. I have co-authored two chapters on standards for the book [Coetzee *et al* 2014]. They are still in draft and have been circulated for comments, see Section 2.2.3.5 for more details.

### 2.2.3.4 The South African Spatial Data Infrastructure (SASDI)

The first attempt to build the **South African Spatial Data Infrastructure (SASDI)** began in 1997, with the establishment of NSIF (then a Sub-Directorate in the Department of Land Affairs). It aimed at establishing the technical and policy framework for enabling unimpeded access to, and utilization of, geospatial data for effective and efficient governance, planning and decision making, through all spheres of government [Cooper & Gavin 2005]. As such, South Africa was then a pioneer in the development of SDIs and as with similar initiatives elsewhere, the focus was on standards development, framing policy and institutional arrangements, and developing a clearing house for geospatial data, for which the key part was capturing and publishing standardized metadata in a catalogue [Cooper & Gavin 2005]. By 29 January 2002, when the last valid metadata record was added to the catalogue, there were about 3000 metadata records available. Unfortunately, by then NSIF was in decline, losing most of its staff over an 18-month period for various reasons. Other than the passing of the SDI Act into law in 2004 and the preparations of draft regulations to support the Act, SDI activities effectively ceased in NSIF and their metadata catalogue was no longer operational [Smit *et al* 2009]. However, with CSI starting to operate in 2010 and with NSIF now starting to have the appropriate resources, mandate and leadership from NGI, the situation should improve significantly within the next year. The metadata catalogue is being revived as the Spatial Metadata Discovery (SMD) [George 2010], built using standards-compliant open-source tools such as Geonetwork [2016]. The delays in getting SASDI creates opportunities for private-sector initiatives, such as that of Kloppers [2014], which also shows the need for SASDI.

### 2.2.3.5 Standards

The primary source for **standards** for geospatial data and services is the relevant technical committee of the International Organization for Standardization, ISO/TC 211, *Geographic information/Geomatics*, for which the local mirror committee at the South African Bureau of Standards is SABS/TC 211, *Geographic information*<sup>5</sup>. I have been active in both since 1998, and I am currently Convenor of Working Group 7, *Information Communities*, of ISO/TC 211 and I was the Chair of SABS/TC 211. From 13 to 18 November 2011, South Africa hosted the 33rd Plenary and related meetings of ISO/TC 211 in Centurion and

---

<sup>5</sup>During 2012, its number was changed from SC 71E, as part of a re-organisation of the SABS's committees.

---

## 2. Spatial data infrastructures and geospatial data

---

27 experts from the local community took the opportunity to participate in the meetings (one of the highest participation rates by local experts at meetings of ISO/TC 211). Based on feedback received at the meetings in Centurion, SABS/TC 211 itself also has a high rate of participation.

As at the end of July 2016, ISO/TC 211 has published 51 International Standards and 13 Technical Specifications. As a mark of the quality of ISO/TC 211's work, on 15 September 2010, ISO presented the committee with the *Lawrence D Eicher Leadership Award* for "recognition of superior performance by an ISO standards development committee that is helping meet the needs of users of standards worldwide" [Tan 2010].

The best-known standard from ISO/TC 211 is ISO 19115:2003, *Geographic information — Metadata*, which has now been superseded by ISO 19115-1:2014, *Geographic information — Metadata — Part 1: Fundamentals*. For more details see Chapter 5, particularly Section 5.8. As well as participating in the development of the ISO/TC 211 standards, SABS/TC 211 has also developed some local standards, as outlined in Section 3.13. One of these, SANS 1883-1:2009, *Geographic information — Addresses, Part 1: Data format of addresses*, led to several projects on addressing within ISO/TC 211. In particular, ISO 19160-1:2015, *Addressing — Part 1: Conceptual model*, has just been published: its Project Leader was a South African, Prof Serena Coetzee of the University of Pretoria.

The nature of standards and standardization is discussed in Section 3.13. Unfortunately, While ISO standards are open, they are not available for free. Even a token cost of R 1.00 for a standard is a barrier to access in many organisations, because of the bureaucracy involved. As a result, the CSI's Standards Subcommittee has initiated negotiations with the SABS for a site license<sup>6</sup> for all the relevant standards for the CSI, to be paid by the Department of Rural Development and Land Reform (DRDLR), as the mother department for the SDI Act. Needless to say, the process is taking a long time, because DRDLR's legal advisors have no relevant experience and because of uncertainties over who the licence will cover (CSI members, sub-committee members, data custodians, etc). Generally, the SABS does such site licences for large tranches of standards (say, 1000 at a time), so identifying the standards to be covered is not actually an issue, as they will get added as and when needed.

The MafA book, *Guidelines of Best Practice for the Acquisition, Storage, Maintenance and Dissemination of Fundamental Geo-Spatial Datasets: Mapping Africa for Africa (MAfA)*, is planned to have 8 parts covering ontology, standards, acquisition, storage, maintenance and dissemination of and for fundamental geo-spatial datasets, organizational issues, and users' perspectives of fundamental geo-spatial datasets [Clarke 2014]. The part on standards has two chapters, *Chapter 10. Standards for the acquisition and maintenance of fundamental geo-spatial datasets*, and *Chapter 11, Standards for dissemination of fundamental geo-spatial datasets*. These chapters cover the types, development of and implementation of standards, and the key standards bodies, specifically ISO, OGC and the International Hydrographic Organization (IHO). The chapters then describe 65 standards to varying extents (depending on their importance), providing implementation benefits and guidelines. These chapters on standards have been circulated for comments [Coetzee *et al*

---

<sup>6</sup>The CSIR has a site licence for all SANS, for example.

## 2. Spatial data infrastructures and geospatial data

---

2014].

The Open Geospatial Consortium (OGC) is an international industry consortium of companies, government agencies and universities (478, as at 28 July 2014, with 2 from Africa), “participating in a consensus process to develop publicly available interface standards” that “support interoperable solutions that ‘geo-enable’ the Web, wireless and location-based services and mainstream IT” [OGC 2016]. ISO/TC 211 and OGC have a cooperative agreement and work closely together, with a number of ISO/TC 211 standards and OGC specifications being identical. In general, ISO/TC 211’s standards tend to be more abstract and OGC’s specifications more directly implementable, but within the framework defined by the ISO/TC 211 standards. Ota & Plews [2015] have used the standards from ISO/TC 211 and OGC as the basis for a software tool for teaching geospatial information technologies and standards.

### 2.2.3.6 SPOT multi-government licence

In April 2007 in South Africa, the first **multi-government licence** anywhere in the world for data from the SPOT 5 satellite came into being. It makes available to all government entities in the country on all three tiers (national, provincial and local), as well as to universities and schools, ortho-rectified<sup>7</sup> and mosaicked<sup>8</sup> images [CSIR 2008]. This agreement was extended recently for the higher-resolution imagery from the SPOT 6 and 7 satellites [EE Publishers 2013].

The SPOT multi-government licence has enabled products such as Eskom’s SPOT Building Count (SBC) since 2006 [Mudau 2010], which now has a library of ten annual inventories of all the buildings in South Africa. In turn, SBC has been used for disaggregating socio-economic data into the mesozones in the CSIR’s Geospatial Analysis Platform (GAP) [Mans 2011].

### 2.2.3.7 Education and training

SASDI needs stakeholders that understand its purpose and benefits, and the value of geospatial data for planning and decision-making. However, as such stakeholders have a wide variety of backgrounds, skills and knowledge, there is a need for **appropriate educating and training** of stakeholders on SDI concepts, which is inadequately catered for by the current GISc academic model of the South African Council for Professional and Technical Surveyors (PLATO) [Rautenbach *et al* 2012b]. Nevertheless, as this study by Rautenbach *et al* [2012b] was conducted with the CSI Sub-committee on Education and Training, the matter is being addressed. Further, Rautenbach & Coetzee [2013] also conducted a survey of available SDI-related education and training material and compiled a database of the results.

---

<sup>7</sup>That is, corrected geometrically from the satellite’s raw imagery to a rectangular grid. The geometry of the raw imagery depends on how the satellite’s sensors work and on the stability of the satellite.

<sup>8</sup>Imagery can be obscured by clouds or have other flaws, so it is common to combine parts of different images together in a mosaic.

---

## 2. Spatial data infrastructures and geospatial data

---

In mid-2013, GISc education was offered at 21 of the then 23 universities in South Africa<sup>9</sup>, but the opportunities are limited for those working full time. Further, these GISc courses relate primarily to *environmental balance* dimension of *sustainable development*<sup>10</sup>, but not to the other three dimensions, *economic growth*, *social inclusion* and *culture* [Coetzee *et al* 2013a]. I would suggest that a national SDI should also support all four of these dimensions.

In July 2014, CSI and GISSA circulated a survey of the demand for GISc knowledge and skills by organisations in South Africa. The intention is to inform the development of SAGIMS to support implementing the NDP.

### 2.3 Geospatial data or geographical information

#### 2.3.1 Overview

Those who deal with geospatial data use diverse terminology and concepts, which can be confusing. Hence, this section contributes to understanding SDIs and VGI by explaining the terminology, types and complexities of geospatial data. In particular, it presents a comprehensive model of a **feature** in Section 2.3.5.

#### 2.3.2 The terminology for geospatial data

As mentioned above, there is a plethora of terms used for *geospatial data*. Thurston [2013] suggests that different terms are more popular in different parts of the world and laments the “splintering and diffusion” that results. The following is a summary of the differences between these terms. Both the hyphenated and un-hyphenated versions of each term are used in the literature.

An attempt was made to ascertain the first time each of these terms was used in the literature by searching for them on Google Scholar [Google 2016e], but it failed because so many of the citations picked up by Google Scholar are incorrect — eg: a search for the term “georeferenced” discovered a paper evaluating spaceborne synthetic aperture radar data that was allegedly written in 1729 [Armenakis *et al* 1729]! The key problem is that search engines are dependent on the quality of the data made available to them on Web sites — in this case, the lists of references in scholarly publications available online. This topic has been explored for various disciplines by many scholars, such as by Wright & Armstrong [2008] for operations research. This issue of the quality of the literature is discussed below in Section 4.8.1.

- *Spatial data or spatial information.*

Of all these terms, this has the broadest scope as it is not necessarily tied to the earth and hence encompasses data of the other celestial bodies. Some feel that the term

---

<sup>9</sup>Sol Plaatje University and the University of Mpumalanga opened in 2014.

<sup>10</sup>According to Agenda 21 [United Nations Conference on Environment and Development 1992].

## 2. Spatial data infrastructures and geospatial data

---

is too broad and that it also includes all geometrical models and Cartesian spaces, which are really in the domain of *computer-aided design* (CAD), and other manifolds.

- *Spatially-referenced data* or *spatially-referenced information*.  
This emphasises that the data or information have a spatial context and hence includes *non-spatial* data or information, and not just coordinates and the like.
- *Geo-spatial data*, *geospatial data*, *geo-spatial information* or *geospatial information*.  
The prefix *geo* means earth, and hence ties the spatial data to the earth. This is possibly the preferred term to use and the one I have used here.
- *Geographical data* or *geographical information*.  
Data or information relating to geography, “the study of the physical features of the earth and its atmosphere, and of human activity as it affects and is affected by these, including the distribution of populations and resources and political and economic activities” [Oxford Dictionaries 2016]. However, while such a definition makes the terms eminently suitable, in my experience there are some who feel that the terms refer only to data and information used by geographers, such as demographic, socio-economic, environmental and land cover data and information, excluding domains such as land administration and engineering, for example.
- *Geographic data* or *geographic information*.  
The version of *geographical* preferred by the Americans as the latter term is deemed to be of “British usage” [Abler 1987]! Ironically, then, the British chose to use it for the title of their umbrella organisation, the *Association for Geographic Information* (AGI), launched in January 1989 [Cooper 1993]. AskOxford, the online version of the Compact Oxford English Dictionary, considers *geographic* to be a derivative of *geographical* [Oxford Dictionaries 2016].
- *Geographically referenced data* or *geographically referenced information*.  
Perhaps the most accurate label, consisting of all data or information that refer to the human-environment system and that can be localized in space and time [Cooper 1987b]. However, it is little used because it is unwieldy.
- *Geo-referenced data* or *georeferenced information*.  
This is the shortened form of *geographically referenced data* or information.
- *Geo-data*, *geodata*, *geo-information* or *geoinformation*.  
These are further shortened forms of *geographical* or *geographically referenced data* or information. The latter are used in the name of the society for representing the community in South Africa, namely the *Geo-Information Society of South Africa* (GISSA), and for the *Geoinformation Subcommittee* of the Committee on Development Information, Science and Technology (CODIST), of the United Nations Economic Commission for Africa (UN ECA).
- *Land data* or *land information*.  
This is an “archaic” term for the data used in a *land information system* (LIS), covering fields such as land administration, cadastre, deeds, property ownership and social tenure, but **not** fields such as demography, economics, natural resources or land cover.

## 2. Spatial data infrastructures and geospatial data

- *Cartographic data or cartographic information.*  
This concerns the rendering of *spatial data* and *non-spatial data* into a map<sup>11</sup>, including aspects such as colours, patterns, textures, line styles, symbols, annotation, multiple languages, label placement, grids, generalization and displacement. These issues are discussed in more detail in Section 2.9. Unfortunately, this term has become ambiguous because of its casual use to describe *spatial data* as well. Strictly speaking, in the cartographic context there are three types of data used to make a map: the raw *spatial data*, the *cartographic data* and the *map furniture*, which surrounds the map, such as scale bars, North arrows, legends and coordinates.
- *Map data or map information.*  
This is an alternative for *cartographic data*.
- *GIS data or GIS information.*  
This is what is used in a *geographical information system (GIS)*. This is the most colloquial of these options, likely to be used in speech or informal writing only.

All of these refer to both digital and analogue<sup>12</sup> data, but we are considering only digital data in this thesis. While much analogue spatial data are created all the time (eg: sketch maps of directions), they do not contribute directly to SDIs. They could be digitised, of course, so that they could be added to an SDI.

The terms “data” and “information” are sometimes also used interchangeably — for example, the term “spatial information” has surely been used when “spatial data” would have been accurate, and vice versa. Technically, information is knowledge that was not previously known to its receiver. The information  $I(x)$  for event  $x$  of probability  $p(x)$  is given by  $I(x) = -\log p(x)$ , that is, the information is highest for the least probable event [Longley & Shain 1982]. Essentially, extracting meaning or structure from data produces information, that is, separating the signal from the noise.

This hierarchy is taken further in information science to include knowledge and wisdom, but these concepts are more subjective and are hence exploited by marketeers. For example, Newdea, a vendor of patented monitoring and evaluating software, states on their Web site that “*we were founded on the principle that the social sector could improve dramatically if it advanced from data and storytelling to becoming information and knowledge driven*” and “*it has been shown that companies and organizations outside of the social sector became more impactful as they advanced up the DIKW continuum*” and “*our customers consider us as the trusted partner that turns their data into information, knowledge and wisdom as they try to change the world*” [Newdea 2016].

### 2.3.3 The types of geospatial data

From the literature review for their survey on core data sets and custodians for the South African Spatial Data Infrastructure, Schwabe & Govender [2012] identified the following terms as being interchangeable in describing a data set: *reference, base, framework, primary,*

<sup>11</sup>Eg: paper, electronic or tactile map.

<sup>12</sup>Hard copy or paper maps.

## 2. Spatial data infrastructures and geospatial data

*fundamental*, *core* and *foundation*. They identified that while these terms might result in similar definitions, they relate to two perspectives that are quite different: the key data providing structure and connections for other data sets, and the important and widely used data sets. However, it does not matter that the terms are not used consistently, as invariably, the data providing structure and connections are also widely used, and the widely used data sets provide structure and connections for other data.

Schwabe & Govender [2012] define core data as a set of geographic information that is necessary for optimal use of most GIS applications, that is, which are a sufficient reference for most geo-located data, and then state that core may refer to the fewest number of features and characteristics required to represent a given data theme.

The respondents to the study identified 119 core geospatial data sets, which could be considered to be themes, as each of them could comprise a variety of types of geospatial data. For example, using some of the concepts described in ISO 19123 [2005], a data set of *crops* could contain a *continuous coverage* (eg: a probability surface derived from a remotely-sensed image, showing for each pixel, the likelihood that it contains the target crop), a *discrete coverage* (eg: wherein each field has an homogeneous coverage of one crop), a *direct position* (eg: the coordinates of where an observation was made of the crop), a *geometry set* (eg: a collection of points, such as observations of a crop), a *solid* (eg: the three-dimensional volume of a fruit-bearing tree) or a *feature attribute* (eg: where the type of crop is an attribute of some other feature type, such as a field). Further, a data set could contain a *linear* feature, with its location defined using a variety of geometric primitives such as line strings, arcs, Bezier curves or clothoids (eg: a trellised vine), or a *polygon* or *area* feature, with its location being the interior of a boundary defined by linear features (eg: a cadastral property that is a farm).

Rautenbach [2011] points out that key to identifying a data set as fundamental is its *value*: while invariably quantitative, it is not necessarily expressed in financial terms. The value could include aspects such as “number of lives saved, improvements in environmental quality, or enhanced regulatory efficiency” [Rautenbach 2011], which depend on the attributes included, quality of the data, and spatial, spectral and temporal resolution. Further, she points out that fundamental data sets need to comply with standards and pass validation processes for quality, etc (see Chapter 6 for a discussion of the dimensions of quality).

A *point of interest (POI)* is a point feature of specific significance, with a direct position, classification, name and possibly other attributes. A POI can also be used to position a data set. The term POI is probably used most often in the context of portable navigation devices and repositories of VGI, where a POI is probably of more interest to a consumer using the data (eg: to find a particular type of government or private-sector service point [Schwabe & Govender 2012]), than to a professional creating data sets. Hence, a data set of POIs often covers a variety of themes and can be application-specific, which means that the underlying thematic data sets separately are more appropriate for an SDI than a generic collection of POIs.

[Schwabe & Govender 2012] suggest that POIs “would be all those features that are not covered elsewhere”, but this is not correct, as a POI could also be included in another data

---

## 2. Spatial data infrastructures and geospatial data

---

set as well. For example, the 1:50 000 national mapping series includes the locations of post offices and police stations, which are also POIs. Depending on the scale of the data or how the data were captured and/or processed, a POI could actually be a polygon, solid or some other type of geospatial data.

### 2.3.4 The complexities of geospatial data

Unfortunately, digital geospatial data are complex, as shown in this chapter. Abstract concepts describing the fundamental nature of digital geospatial data have to be made concrete in a GIS so that they can be rendered to data structures and code. Various spatial models are described below in Section 2.5. Much of the complexity of a GIS lies in the complexity of geospatial data, such as the need to represent infinite point sets through approximation, using line segments instead of high-order polynomials and discrete instead of continuous spaces [Rigaux *et al* 2002]. Unfortunately, because of the ready availability and apparent ease of using a GIS, virtual globe or global navigation satellite systems (GNSS<sup>13</sup>) receiver, many inexperienced users are unaware of these complexities (especially coordinate reference systems).

All digital data *attempt* to model and describe the world, for computer analysis, display, inventory, etc. Digital data are always only an abstraction of reality; they are always partial (reflecting the conscious and unconscious biases of the compilers); and they are not complete. Any set of digital data is always only just one of many possible ‘views’ of the world — they cannot be an exact duplication of the world. Such attempts at maps with perfect representations of the real world have been parodied in literature by the likes of Carroll [1892], with *Mein Herr’s* map that would block out the sun, and Borges [1946], with his *Map of the Empire*, both at a scale of 1:1, and in the cinema by the likes of *Synecdoche, New York*, with its 3D recreation of New York [Kaufman 2008].

For any data set, some things are approximated, some things are simplified and some things are ignored. Hence, there can never be perfect, complete and correct data! So, to ensure that data are not misused, all the assumptions made in creating a data set and all of the limitations should be documented fully (ie: as *metadata*)<sup>14</sup>, as discussed below in Sections 2.7 and 5.2.

Digital geospatial data provide a digital representation of part of the real and/or potential world.

- **Real world.**

The world as it is or as it was. This would include current records, as well as historical data. The historical data would be used for time series analysis (studying the changes in the human-environment system) or for archaeological or historical purposes.

---

<sup>13</sup>The United States’ NAVSTAR global positioning system (GPS) is the best-known GNSS and the only one that is fully operational on a global scale, but Russia’s GLONASS is close to full operation, China’s Beidou (available for civilian use from January 2013 [StrategyPage.com 2013]) and France’s DORIS are operational on a regional scale, and systems are under development by the European Union, Japan and India.

<sup>14</sup>The introduction to ISO 19115 [2003] contains similar wording, because I contributed some of the text.

## 2. Spatial data infrastructures and geospatial data

- **Potential world.**

The world as it might be or might have been. This would include forecasts, simulations and scenarios, and would be used for planning purposes, or for filling “gaps” in the historical data, for example [Cooper 1993].

### 2.3.5 Feature

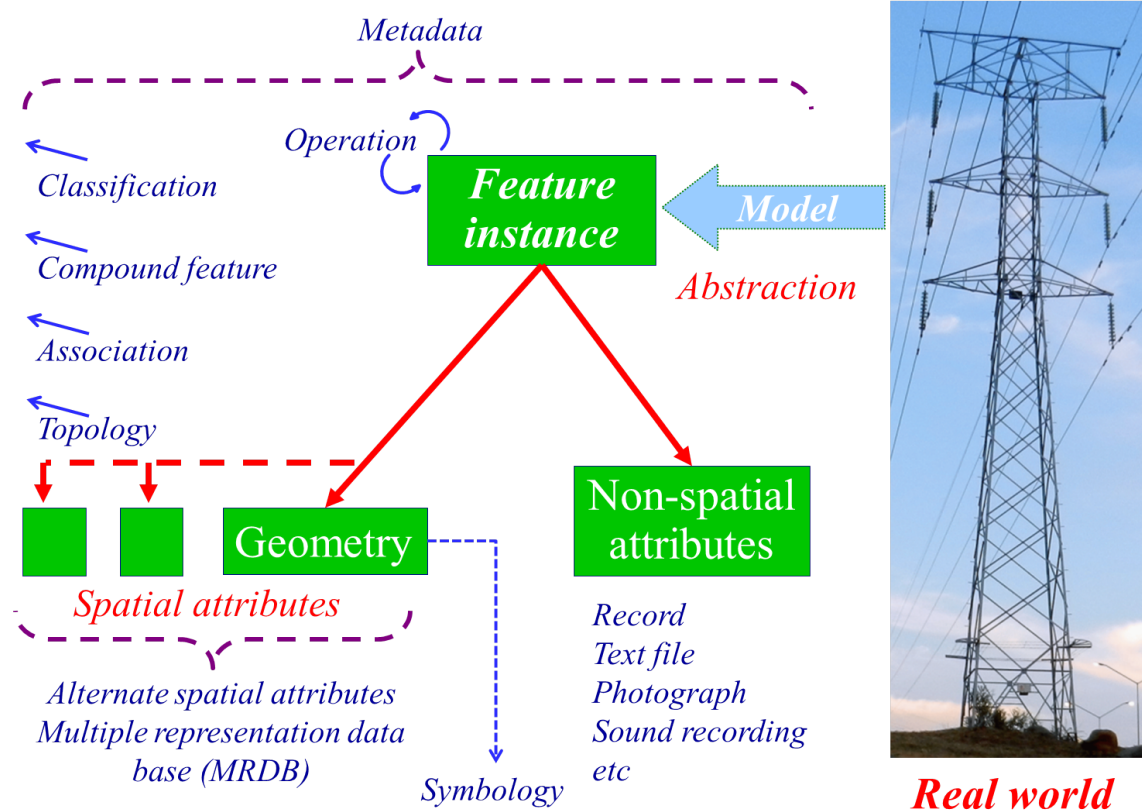


Figure 2.2: A feature and related concepts

Figure 2.2 provides a conceptual model of geospatial data. When a person views the world, they don’t see geospatial data. Rather they see things (*phenomena*) that are *represented* by geospatial data (for example, the pylon shown in Figure 2.2). These things are then *modelled* in a GIS as *abstract concepts*, where the preferred term for the abstract representation of a thing is a *feature* (also known as a geographical *entity* or *object* [Rigaux *et al* 2002]). One exemplar of a feature is an *instance* and features are grouped (*classified*) as *feature types* or *classes*. Note that as the word *class* carries so many different meanings in different contexts, the preferred term is *feature type*. A *compound feature* or *feature concept* can consist of other features of various types. The *spatial attributes* (the *geometry* or *coordinates*) and the *non-spatial attributes* (or *feature attributes*) are then attached to the feature. Figure 2.2 illustrates the concepts related to a feature and Table 2.1 describes them and provides examples. These concepts are also described in more detail in the following

## 2. Spatial data infrastructures and geospatial data

sub-sections. *Classification*, *association* and *topology* all connect a feature to other features, as represented by the arrows in the figure pointing outwards.

There is a correspondence between this *feature model* and *formal concept analysis (FCA)* [DG Kourie 2014, *pers comm*]. This is discussed briefly in Section 9.3.


Considering features in terms of the terminology of an object-oriented paradigm, *complex objects* are implemented through feature concepts or compound features. Each feature instance has a unique *object identifier* which ensures its existence independently of its value, facilitating sharing attributes (eg: mutual boundaries) and updating (eg: correcting a positional error in a mutual boundary). Further, *types* and *classes* are implemented as *feature types* and *inheritance* as feature sub-types (with hierarchical classification), though generally none of them with *methods* [Atkinson *et al* 1989; Rigaux *et al* 2002]. For more on feature-based and object-oriented GIS, see Sections 2.5.3 and 2.5.4 respectively.

Table 2.1: Features and related concepts

Concept	Description	Example
Feature	An abstraction of real world phenomena, as a <i>feature concept</i> , <i>feature type</i> or a <i>feature instance</i> [ISO 19101-1 2014]	
Feature concept (compound feature)	An aggregation of one or more feature types	A nature reserve consisting of a boundary fence, entrance gate, camping site, public roads, restricted areas, etc
Feature type (class)	A logical grouping of feature instances, based on their common characteristics	nature reserve
Feature instance	A discrete phenomenon in the real (or imaginary) world represented in a data set Something specific out there that is modelled in the data set. The instance normally has coordinates and may be portrayed on a map by a particular graphic symbol	Nylsvoley Nature Reserve
Non-spatial attribute	A characteristic of a feature. Has a name, a data type and a value domain, A feature attribute may occur as a type or an instance	Important Bird Area (IBA) Category
Attribute value domain	The set of valid values for an attribute	real number
Attribute value	The value for a particular attribute of a particular feature	Global IBA (A4i, iii). > 1% of a species, > 20 000 waterbirds
Association	A relationship linking one feature to another	Linking a catchment to a river

Continued on next page

## 2. Spatial data infrastructures and geospatial data

Concept	Description	Example
Operation	An action that can be performed on a feature	<i>Upgrading an IBA Category</i>
Spatial attribute	The spatial representation of a feature, including geometry, topology and alternate spatial attributes	
Geometry	The actual coordinates for a spatial attribute	24° 37' 30" S 28° 40' 30" E
Topology	The spatial relationships between the geometry of features that are invariant to continuous transformation of the geometry	The Nyl River intersects with the Nylsvley Nature Reserve
Alternate spatial attribute	One feature instance can have several spatial attributes, for representation at different scales or for different types of spatial analysis. This concept is also known as a multiple representation database (MRDB)	Nylsvley as a point feature at 1:1M and as an area feature at 1:50K
Symbology	A graphic representation of a feature, typically using the feature's geometry for its position and shape, and the feature's type and attributes for its colour, etc	
Metadata	Descriptions of all the concepts	The data are what I remember being told as a Friend of Nylsvley and the Nyl Floodplain

### 2.3.6 Geometry and topology

Typically, the location of digital spatial data in space and time is recorded in two or even three spatial dimensions — only rarely is its location recorded in the temporal dimension [Cooper 1987b], even now. While there used to be hybrid models (eg: the *vaster* format proposed by Peuquet [1983]), spatial data sets are now either raster or vector.

- **Raster.**

This is a disassembly of continuous (or even discontinuous and/or overlapping) space, with the data stored as a tessellation (typically rectangular), with one or more values associated with each element (cell). Such data are also known as being field- or space-based. The more generic term is a *grid*, a “network composed of two or more sets of curves in which the members of each set intersect the members of the other sets in an algorithmic way” where “the curves partition a space into grid cells” [ISO 19123 2005]. The cells could be of any shape and the grid could be regular or irregular, but in practice, raster data are usually implemented as a regular grid.

---

## 2. Spatial data infrastructures and geospatial data

---

- **Vector.**

This is the construction of data structures from the data, with the data stored as a set of nodes, lines, curves, areas, surfaces and/or solids having position and as appropriate, magnitude and direction [Rigaux *et al* 2002; ISO 19123 2005].

To some extent, raster data could be viewed as a top-down sub-division of an area, while vector data could be viewed as a bottom-up identification and aggregation of phenomena in the area.

Most GISs provide some integration of both vector and raster data together, primarily for overlay on one another (eg: for 3D perspective views, fly-throughs or viewshed analysis), but also for deriving new data (by inspection or analysis), detecting errors, re-projecting data sets or propagating updates across data sets [Butenuth *et al* 2007]. Both forms have their advantages: the vector form is efficient for storing data, for catering for multiple representations (or alternate spatial attributes, see 2.1) and for performing network analysis, while the raster form is efficient for remotely-sensed imagery and performing polygon overlays, for example.

It is possible to convert vector data into raster data and vice versa, though generally the former conversion is easier than the latter [Cooper 1993]. Rigaux *et al* [2002] suggests that a **feature** cannot have raster or field-based spatial attributes, but this is not correct. For example, ISO 19123 [2005] defines a *coverage* as a **feature** *that acts as a function to return values from its range for any direct position within its spatial, temporal, or spatio-temporal domain*.

The *topology* of geospatial data is based on mathematical topology, being the spatial relationships between the geometries of features that are unaffected by changes to the shape or size of spatial attributes. That is, the spatial relationships are invariant to transformations of the geometry, such as projection from an ellipsoid to a plane. Topology provides information on the spatial relationships inherent in the data, such as:

- **Coincidence:** where more than one feature shares the same spatial attributes;
- **Intersection:** a special case of coincidence, where two or more lines cross or meet at one point;
- **Containment:** where a feature lies wholly within another feature;
- **Inclusion:** a type of containment, where the included feature forms part of the containing feature;
- **Exclusion:** a type of containment, where the included feature does not form part of the containing feature (also known as an *island*); and
- **Adjacency:** where areas share a common boundary and lie on opposite sides of the boundary [Cooper 1993].

Note that *inclusion* and *exclusion* are dependent on the semantics of the data and not just on the geometry and topology, so they have to be identified explicitly. They cannot be determined automatically from the spatial data, but might be inferred by an expert system through using the non-spatial data as well.

## 2. Spatial data infrastructures and geospatial data

---

The visual superimposition of data sets can often reveal their shared lineage through inheritance of the same positional errors [Goodchild 2008b]. Similarly, such visual superimposition can allow a human to detect topological relationships in the data. However, while some forms of topology can be inferred or calculated from the data (eg: by identifying the features that share spatial attributes), others need to be stored explicitly (eg: exclusion). A topological map allows one to emphasise the structure of data rather than the exact locations. Perhaps the most famous topological map is that of the London Underground issued by London Regional Transport, which has been adapted around the world to show other railway networks and bus routes [Cooper 1993].

### 2.3.7 Non-spatial data

*Non-spatial attributes* are independent of the position of the feature, that is, they are all the data stored about a feature, excluding the spatial data. They describe the nature and appearance of the feature, such as its name, function, capacity, composition, owner or colour. As with a feature, a non-spatial attribute may occur as a type or as an instance. Typically, a non-spatial attribute has a name, a definition, an encoding (for compactness in a database), a data type and a value domain.

The *non-spatial* component of digital geospatial data has many appellations, though they all refer to the same thing: alphanumeric data, descriptive data, text, attribute, non-spatial data, multi-media, etc. These are *characteristics* of features and are related to each other through features. Further, an *association* and an *operation* are special characteristics of a feature.

- **Association:**  
This is a relationship linking one feature to one or more other features, such as the association between a river and its catchment area, or the relationship between a bridge and both what it carries and what it spans.
- **Operation:**  
This is either an action that can be performed on a feature, such as upgrading or downgrading an attribute (eg: changing the category for an Important Bird Area), or a query that can be performed on a feature, such as checking the water level at a weir. However, as operations are for processing data, they are not often included in data sets being exchanged.

As discussed in Section 2.4.1, there is both a duality and a grey area between the *non-spatial attributes* of a feature and the *classification* of a feature [Cooper 1987a].

## 2.4 Classification, taxonomy, ontologies, folksonomies, etc

Chapters 8 and 9 present core research findings in assessing various repositories of volunteered geographical information against several taxonomies of user-generated content. This section discusses the nature of classification (which goes under a variety of names),

---

## 2. Spatial data infrastructures and geospatial data

---

which is often done incorrectly. Typical problems are presented in Sections 2.4.5 and 2.4.6. A folksonomy is a user-generated taxonomy and an ontology is essentially a taxonomy with inference (or implicit relationships), and both are also discussed in this section.

Some consider classification to be part of metadata (see Chapter 5), such as Duval *et al* [2002]; HLWIKI Canada [2015c]. This is obviously not the same as classifying metadata into types of metadata, as standards for metadata do, see Section 5.8; nor the same as the metadata of classification (ie: documenting the classification system used); nor the same as using metadata to classify data sets or other resources and things, eg: Web sites [Pierre 2001] and Twitter users [Nagpal & Singhal 2014].

### 2.4.1 The nature of classification

“To classify, in the primitive sense, is to divide existents of the universe of discourse — concrete or conceptual, things or ideas — into groups” [Ranganathan 1951]. Classification helps with managing large amounts of information, but also with describing things (illuminating a field of knowledge) and then even being able to predict aspects of things that have been classified: classification can even be prophetic [Ranganathan 1951]. This is because different features in the same class share similar attributes. The second sense of classification is to arrange the resulting classes in a preferred order and the third sense is to encode the classification [Ranganathan 1951] (though he limited it to assigning numbers to classes). See Section 2.4.6 for a discussion on encoding a classification.

Ranganathan devised the *Acknowledgement of Duplication*, whereby having any one system of classification of information necessarily implies that there is at least one more, and different, classification for any thing in the classification. He then introduced *faceted classification* through his *colon classification system*, which allows things to be classified through five facets or perspectives that supplement the base hierarchical classification: *Personality* (distinguishing characteristic), *Matter* or *Property* (physical material), *Energy* (actions), *Space* and *Time*. This enables a classification scheme to be flexible, rather than just ready-made [Ranganathan 1951].

The classification of information is a subjective process because people observe different properties in the things being classified and require information about the things to different levels of detail [Scheepers *et al* 1986]. Classification is also a social and political construct. Consider, for example, the different perspectives of the policy maker, the technocrat, the business person, the subsistence farmer, the consumer, the researcher, etc. This correlates with Ranganathan’s *Acknowledgement of Duplication*, as well as the principle of *complementarity* [Bohr 1937], whereby different views represent different aspects of the world, each of which may be appropriate in some context. While they might appear to be incompatible, they are also essential for an exhaustive understanding and accentuate different aspects [Bohr 1937; Kokla & Kavouras 2001]. For the same reasons, there is both a duality and a grey area between the classification of features and the non-spatial attributes of features [Cooper 1987a]. The duality is that a feature’s class (feature type) could itself be viewed as a non-spatial attribute of the feature. But it is also a grey area because deciding between using a class or an attribute to record the characteristic is open

## 2. Spatial data infrastructures and geospatial data

---

to interpretation, depending on one's application and perspective, both of which change (the selection is fuzzy); the characteristic could be recorded simultaneously in both the class and in one or more attributes; and the relevant characteristics could be ignored.

### 2.4.2 Terms for classification

There are many different structures that can be used for a classification, which I have discussed before in Cooper [1993]. The structure can introduce relationships between classes (or types), such as super- and sub-classes, and the inheritance of attributes. Each class needs to be defined and labelled, and is normally given a code (see Section 2.4.6). There is also a variety of words used for classification:

- **Classification:** "the action or process of classifying something ... (biology) the arrangement of animals and plants in taxonomic groups according to their observed similarities ... a category into which something is put" [Oxford Dictionaries 2016];
- **Catalogue:** "a complete list of items, typically one in alphabetical or other systematic order" [Oxford Dictionaries 2016];
- **Categorization:** "placement in a particular class or group (adapted from the definition of *categorize*") [Oxford Dictionaries 2016];
- **Codification:** "arrangement (of laws or rules) into a systematic code ... arrangement according to a plan or system (adapted from the definition of *codify*" [Oxford Dictionaries 2016];
- **Glossary:** "an alphabetical list of words relating to a specific subject, text, or dialect, with explanations; a brief dictionary" [Oxford Dictionaries 2016];
- **Nomenclature:** "the devising or choosing of names for things, especially in a science or other discipline ... the body or system of names used in a particular specialist field" [Oxford Dictionaries 2016];
- **Taxonomy:** "the branch of science concerned with classification, especially of organisms; systematics ... a scheme of classification" [Oxford Dictionaries 2016];
- **Terminology:** "the body of terms used with a particular technical application in a subject of study, theory, profession, etc" [Oxford Dictionaries 2016];
- **Thesaurus:** "a book that lists words in groups of synonyms and related concepts" [Oxford Dictionaries 2016];
- **Typology:** "a classification according to general type, especially in archaeology, psychology, or the social sciences" [Oxford Dictionaries 2016]; and
- **Folksonomy:** a neologism and portmanteau of *folk* and *taxonomy* meaning collaborative tagging, or the classification and identification of content by the general public, rather than by domain experts. A folksonomy is uncontrolled, which is both its strength and its weakness. Vander Wal [2007]<sup>15</sup> defines folksonomy as "tagging that

---

<sup>15</sup>He is credited with inventing the neologism, but on a private mailing list.

---

## 2. Spatial data infrastructures and geospatial data

---

works” and that it had three tenets: a *tag*, the *object* being tagged and an *identity*, but see Section 2.4.3 for a fuller discussion of the concept.

Given the similarities in the definitions of *classification*, *catalogue*, *categorization*, *codification*, *taxonomy* and *typology*, I have chosen to use the word *taxonomy* here, because it is the term most closely associated with scientific classification. Five taxonomies of user-generated content, from Wunsch-Vincent & Vickery [2007], Gervais [2009], Budhathoki *et al* [2009], Coleman *et al* [2009] and Castelein *et al* [2010], and one of citizen science Wiggins & Crowston [2011], are assessed in Chapters 8 and 9.

Hopefully, a taxonomy should comprise classes that are *logical* groupings based on the *common characteristics* of the things being classified. It should be robust, such as being defensible in court [Cooper 2003]. Otherwise, one could end up with something like the classification of animals in the “Celestial Emporium of Benevolent Knowledge” by Borges [1952]:

*“1) Belonging to the Emperor, 2) Embalmed, 3) Trained (or Tamed), 4) Sucking pigs, 5) Sirens, 6) Fabulous, 7) Unleashed dogs (or Stray Dogs), 8) Included in this classification (or Included in the present classification), 9) Which jump about like lunatics (or Frenzied), 10) Innumerable, 11) Drawn with a fine camel-hair brush (or Drawn with a very fine camel-hair brush), 12) Et cetera, 13) Which have just broken the pitcher (or Having just broken the water pitcher), 14) Which look from a distance like flies (or That from a long way off look like flies)”.*

A naïve, automated classification of big data could well produce something similar to this parody from Borges [1952], unless it is reviewed critically.

### 2.4.3 Folksonomy

A traditional classification respects Aristotelian contraries, while a folksonomy takes a non-Aristotelian approach, allowing tags that some might consider to be true but others might consider to be false. Further, folksonomies do not necessarily deal well with typographical, spelling and other errors and variations (eg: harbour vs harbor; slip road vs ramp; and school vs college vs university), so some consider them to be unsophisticated [Peterson 2006]. Within the OpenStreetMap community, for example, those who contribute the most do not necessarily participate in the voting on tags, creating tensions [Perkins 2013].

*“Folksonomy is the result of personal free tagging of information and objects ...for one’s own retrieval”* and is *“created from the act of tagging by the person consuming the information”* [Vander Wal 2007]. Hence, here the folksonomy is the classification as *perceived by the user* (so it can only exist if shared with others), which is not necessarily the same classification perceived by the creator of the tag at the time of the tagging. Further, he limits a folksonomy to tagging in a *social environment* and to the tagging of objects with *universal resource locators (URLs)* [Vander Wal 2007]. It is unlikely that these latter restrictions are widely known and I contend that they are not necessary. I was not aware of them, for example, until I studied the text of Vander Wal [2007].

## 2. Spatial data infrastructures and geospatial data

---

However, I do agree that “*the value in this external tagging is derived from people using their own vocabulary and adding explicit meaning, which may come from inferred understanding of the information/object*” [Vander Wal 2007]. This allows folksonomies to be dynamic, catering for the rapid changes in technology and explosion in the availability of content (*Facebook* users add about one billion photographs every three days! [Metz 2013]), though it can obviously make some tags unintelligible. Social tagging is also a way of making one’s mark in the Web and to improve one’s unstructured data management, as tags facilitate searching [HLWIKI Canada 2015c]. Social cataloguing sites aggregate tags, allow them to be ranked or weighted, and facilitates interactive dialogue on content and repositories: supporting the Third Law of Ranganathan [1931]: every reader their book [HLWIKI Canada 2015b].

The tagging systems make it possible for a user to see and use the tags created by others, with the resulting common vocabulary becoming biased towards the more popular tags and clusters of tags (*tag clouds*) for resources, rather than the personally-oriented tags. These common or popular tags tend to be accurate representations of the resources [Golder & Huberman 2006; Kalantari *et al* 2010]. Of course, this power law distribution of tag popularity is susceptible to the fallacy in classical logic of *proof by assertion* [Keeler 2011] or *argumentum ad populum*, namely that just because something is popular, it must be true. This discussion of proof by repeated assertion is expanded on in Section 4.3.4.12.

So, “*people are not so much categorizing, as providing a means to connect items (placing hooks) to provide their meaning in their own understanding*” [Vander Wal 2007]. On the other hand, Kalantari *et al* [2010] do not use folksonomies for classifying spatial data, but view them as sources for the automated enrichment of metadata. These different uses of folksonomies do emphasize how easily a new concept or neologism can be misunderstood, and possibly even misused! Further, tagging can effectively also be used for making recommendations through the nature of the tag (eg: *restaurant* vs *excellent restaurant*).

Tagging or folksonomies is one aspect of *social annotation*, which also includes users highlighting extracts in text books and making notes in them. Gazan [2008] suggests that these benefit subsequent users of the text books, even where the source and rationale for the annotation is unknown. Hence, he would like to see social annotations of digital library collections. As with folksonomies, these would then be readily available to anyone and available for searching and the like, unlike the written annotations in a paper textbook.

### 2.4.4 Ontology

Strictly speaking, there is only one *ontology*, which is “*the branch of metaphysics dealing with the nature of being*” [Oxford Dictionaries 2016]. One aspect (of many!) of ontology in philosophy is to classify things — for assessing existence, essence, qualities, etc.

Unfortunately, *ontology* is now a concept much abused in computer science, now meaning a formal representation of knowledge: the concepts within a domain, the relationships between those concepts, a shared vocabulary and reasoning about the encoded knowledge. Hence, Oxford Dictionaries [2016] now also defines *ontology* as “*a set of concepts and*

## 2. Spatial data infrastructures and geospatial data

*categories in a subject area or domain that shows their properties and the relations between them*". There is then a wide range of interpretations of ontology, such as:

- Standard set of terms and definitions, that is, a fixed and controlled vocabulary (a dictionary or a glossary);
- Taxonomy or classification;
- Extracting implicit relationships (automated reasoning or inference), such as the calculated topology in spatial data;
- As suggested by Smith [2010], the combination of a *taxonomy* with *relationships*; or
- A way to describe different types of pizza!<sup>16</sup>

Adams [2002] suggests that the traditional librarianship skills of "thesaurus construction, metadata design, and information organization" fit in well with the semantic Web. An ontology can be implemented as types, properties and relationship types; it can be controlled by restrictions and rules; it can be assessed through axioms; and it can be modified by events [Yeung & Hall 2007]. There is also a wide variety of formal languages for encoding the information about a domain and, usually, for encoding reasoning rules for extracting the implicit relationships in the information. Examples of these languages include OWL (Web Ontology Language), DAML (DARPA Agent Markup Language), OIL (Ontology Inference Layer), IDEF5 (Integrated Definition for Ontology Description Capture Method), DOGMA (Developing Ontology-Grounded Methods and Applications) and CASL (Common Algebraic Specification Language). Ontologies are also used in the *Semantic Web*, see Section 3.5.

### 2.4.5 The dangers of classification

*"Classification and social conventions allow us to broaden the network of social relationships by making network of networks, and this in turn allows us to create very large groups indeed. Of course, the level of the relationships is necessarily rather crude but at least it allows us to avoid major social faux pas at the more superficial levels of interaction when we first meet someone we don't know personally"* [Dunbar 2010, p 80]. Hence, the need for stereotypes!

#### 2.4.5.1 Dangers identified by Shirky

A taxonomy not only provides *description*, it also provides *prediction*, because each class has certain characteristics that one can expect of new items placed in the class [Shirky 2005], that is, stereotypes. However, Shirky [2005] also identified a number of problems with classification systems.

- The creator of the classification could be unaware of the *constraints* of the context in which a taxonomy is defined, which can create errors in the taxonomy, such as the incorrect characteristics being used to differentiate classes.

<sup>16</sup>It seems that all training materials on ontologies use pizzas for examples.

## 2. Spatial data infrastructures and geospatial data

---

- Effectively, to avoid such constraints might require the creator to do *forecasting* or *fortune-telling*, to ensure that the taxonomy remains robust as the context or environment changes.
- It is incorrect to impose a *single 'authoritative' hierarchy*, rather than allowing *partially-ordered sets (posets)*, where one class can inherit from several 'parent' classes, or allowing *hyper-links* between classes.
- A *voodoo categorization* is an expectation that naming the world (or putting things in classes) changes it.
- *Similar terms might actually encode different things* and it might be better to separate, not combine, them — eg: the labels 'cinema', 'film', 'movie' and 'flick' are similar but might identify classes of people with different, though perhaps overlapping, interests in motion pictures who would rather be separated than combined, to filter out the films not of interest to them.
- The upper levels of a hierarchy can actually be *unstable*, eg: defunct countries.
- Finally, can an expert hierarchical taxonomy provide more certainty than raw *searching*? I would suggest that mere searching is the equivalent of a naïve classification that does not necessarily use correct characteristics to identify the classes, that is, that fails to address some of the problems described in this section. A good classification should provide more than just the ability to find something, such as providing the ability to predict characteristics sensibly [Ranganathan 1951]. This 'debate' over searching *vs* classification is similar to that over metadata *vs* searching, see Section 5.10.

### 2.4.5.2 Dangers I have identified

Other problems that I have previously identified with a classification system [Cooper 2003; Cooper & Das 2009] are given below.

- It is essential to provide unambiguous and unique *definitions* for classes. It is not sufficient to just have a label (or term) to identify a category: one needs to understand that the category reflects an abstract concept (latent variable), and the label merely identifies the category uniquely. In addition, one also needs a formal definition of the category, which should be readily understandable.
- However, as a definition is not necessarily sufficient to select the category correctly, Bevel & Gardner [2008] consider it essential to have a *category selection mechanism* as well, such as a decision map or a decision tree.
- The *quantitative attributes* can change easily, so they should not be used for differentiating categories. Unfortunately, this is the standard practice with naïve automated and statistically-based classification, which then needs to be tempered by using qualitative characteristics as well.

## 2. Spatial data infrastructures and geospatial data

- *Overloading* a category occurs by mixing up different types of characteristics (such as objective and subjective ones) to identify a category. This results in a hybrid classification in which it is often difficult to place data or add new categories. This occurs when one category is used to convey several different meanings, even though they are often independent (see also Section 2.4.6). An example from bloodstain pattern analysis (BPA) would be basing a category on both the cause and geometry of a bloodstain [Cooper 2003]. This is supported by a key finding of a recent study on the reliability of current methods in BPA that “it would also be advantageous for the BPA community to agree on a standard methodology for the analysis of bloodstain patterns which includes a better distinction between classification and reconstruction and relies less on mechanistic descriptions of patterns” [Laber et al 2014]. A widely used classification with this problem is the International Standard Classification of Occupations (ISCO) [International Labour Organization 2016], which mixes up three groupings of characteristics: themes (the type of work), experience and/or qualifications, and the managerial role, see Table 2.2.
- As per Ranganathan’s *Acknowledgement of Duplication* (see Section 2.4.1), it is not correct to assume that there is only one, unique classification. Invariably, there are likely to be several different ‘views’ of the data.
- It is not possible to develop the *perfect taxonomy* in a committee! Attempting this typically results in ‘analysis paralysis’ and taking too long to complete the work. Generally, the shortcomings of any classification are revealed quickly in the field and one must expect to revise or update any classification regularly.
- There appears to be a fixation with having a *round number of classes*! This is often caused by making the categorization based on some coding scheme, when it should obviously be the other way round. A common example is using a single digit for a level in the hierarchy, giving ten categories, as has been done with ISCO [International Labour Organization 2016], see Table 2.2. See also Section 2.4.6 for a discussion on encoding a taxonomy.

Table 2.2: ISCO-88 (the old version) vs ISCO-08 (the new version) [International Labour Organization 2016]

ISCO-88	Cook was classified in Unit Group 5122	
	[5]	Service workers and shop and market sales workers
	[51]	Personal and protective services workers
	[512]	Housekeeping and restaurant services workers
ISCO-08	Chef has been split off to Unit Group 3435	
	[3]	Technicians and associate professionals
	[34]	Legal, social, cultural and related associate professionals
	[343]	Artistic, cultural and culinary associate professionals

## 2. Spatial data infrastructures and geospatial data

---

### 2.4.5.3 Further classification problems

There are further problems that could arise with classification, as discussed below.

- It is inappropriate to *expect users to understand* how things have been organised, without catering for colloquial terms, synonyms, multiple languages or search terms [EJO Gavin 18 April 2006, *pers comm*].
- It is futile to try to impose a *rigid geographical hierarchy*, because boundaries are likely to change. It is equally futile to try to impose rigid definitions for the likes of *urban* and *rural*, rather than providing enough information to enable users to make their own categorizations [EJO Gavin 18 April 2006, *pers comm*]. One attempt to address this *mappable areal unit problem (MAUP)* [Raper *et al* 1992] is the use of *mesozones* for aggregating and disaggregating data [Naudé *et al* 2008].
- Care must be taken to prevent *pigeonholing*, that is, assigning something to an overly restrictive category [Oxford Dictionaries 2016].
- There will always be *bias* in any classification system, if only because of the ‘limited’ knowledge of those designing the system. An example is how a classification system is structured, as the selection of the characteristics for the top levels can bias perceptions about the classes [Doctorow 2001].
- Complications can arise when using a taxonomy for purposes for which it was *not designed*. This is very common because of the convenience of appropriating an existing taxonomy.

As discussed above in Section 2.4.1, the classification of information is a subjective process. A taxonomy could also be considered to be *metadata* for the content being classified. However, while some of this metadata might be recorded in any event by the creator of the content (eg: when the data were contributed by a domain expert), this will not always be the case — particularly when the metadata might cast the content in a negative light (eg: data contributed for malevolent reasons). If appropriate, parts or all of these taxonomies could be added to a metadata standard, such as ISO 19115-1 [2014].

### 2.4.6 The curse of clever codes

As explained above in Sections 2.4.1 and 2.4.5, any taxonomy should consist of classes or types (which are abstract concepts) grouped logically, based on their common characteristics. Each type needs to be identified by a unique label and more importantly, by an unambiguous and unique definition for the type. In addition, it is convenient to assign a terse code (typically, numeric or alphanumeric) to each type for more efficient storage and processing. Similarly, it is convenient to assign a terse code to each instance of a type.

Unfortunately, encoding is often done incorrectly, by trying to make the codes too clever. Codes should really only be for computer use and the system’s user interface should present the label of the type or name of the instance to the user, not the code. However, it is very common for people to like using codes: I have been birding with several excellent

---

## 2. Spatial data infrastructures and geospatial data

---

birders who record their sightings using *Roberts numbers*<sup>17</sup>, which they know off by heart (there are over 900 of them!), rather than species names.

Essentially, the only intelligence that should be embedded in a code is error detection (such as a parity bit or check digit), possibly also with error correction, to detect and even fix transcription errors. The standard ISO/IEC 7064:2003, *Information technology — Security techniques — Check character systems*, does not cater for error correction, but defines a set of check character systems that can detect all single substitution errors, and all or nearly all single (local) transposition and shift errors, amongst others. The current draft of SANS 1876, *Feature instance identification standard*, uses ISO/IEC 7064.

It is very tempting to include metadata in codes as well, such as a geographical context, authority responsible for the class, ownership, quantities or dates. However, such aspects are prone to change, rendering parts of the code redundant or even confusing. For example, within the South African National Identification Number, much of the capacity of the 11th and 12th digits has been lost because they were used previously to designate race and citizenship. On the positive side, it does include a check digit (the 13th digit). When metadata are included in an encoding, it should be the exception and not the rule, and the designer of the encoding needs to be able to justify the inclusion.

## 2.5 Models for geospatial data in a GIS

A geographical information system (GIS) is a computer-based system that efficiently captures, stores, retrieves, maintains, validates, integrates, manages, manipulates, analyses and displays digital geographically referenced information [Cooper 1993]. The data models used in GISs have evolved and matured as the theoretical understanding of the fundamental nature of spatial data developed, and as processing power and data storage have increased. Unfortunately, this has resulted in archaic concepts being made concrete in a GIS in a manner that not only affects the use of the GIS, but also constrains the way users perceive digital geospatial information. As a result, obscure or obsolete features and primitives are present in the architectures of some GISs. The original architecture of these GISs reflected the state of the understanding of their designers at the time they were first designed, but their architecture has been modified and enlarged over the years to accommodate the changes. A result is anomalies in the way they are used — and even in the way GIS users understand their data, as discussed below. As SDIs are built on the technologies of GISs, these concepts and perceptions have permeated into SDIs, so this section helps with understanding SDIs.

### 2.5.1 Centroids as surrogates for polygons

One example concerns the treatment of *polygons* or *areas* in a GIS. Because a polygon consists of a perimeter and the inferred enclosed space within the perimeter, and because

---

<sup>17</sup>The sequence numbers for the species in the 6th edition of *Roberts birds of southern Africa* [Maclean & Roberts 1985], which then changed for the 7th edition [Hockey *et al* 2005].

## 2. Spatial data infrastructures and geospatial data

---

the perimeter invariably consists of many short line segments or arcs, it was difficult to know how to connect anything to the polygon, such as attributes. The designers could not point to the enclosed space, because it is abstract and does not exist itself in the spatial database. It was also not feasible to point to each and every one of the line segments in the perimeter because of the obvious overhead in both storage and processing speed, and it was dangerous to point to only one line segment in the perimeter, because that could easily be deleted when editing the polygon. Hence, the solution that was adopted was to use the *centroid* of the polygon (the centre of gravity of the polygon) as a surrogate for the polygon. Of course, with concave polygons there is a risk the centroid can lie outside the polygon and with *islands* in the polygon, there is a risk the centroid could lie within one of the islands. Both problems can be addressed by moving the “centroid” to a point within the polygon. One well-known GIS then treated polygons and points as being the same (with points being “degenerated” polygons!). Of course, it is far easier to use a *feature* instead of a polygon, to which the spatial and non-spatial attributes are attached, see Section 2.3.5 above.

### 2.5.2 Layer-based geographical information systems

The first GISs with structure grouped similar data together in *layers*. A GIS was effectively the same as a *computer-aided design* (CAD) system used for designing electronic circuit boards, which would allow for electronic components, such as microprocessors and connectors, to be laid out in layers, as that is how circuit boards were typically constructed then [Cooper 1993]. Indeed, even now, decades later, the basic digitising of geographical data could be done using a CAD and one still sees recruitment advertisements for “CAD operators” to digitise geographical data.

Each layer in the GIS could be limited to a specific type of geometry (eg: points, lines or areas) and/or to a specific type of data (eg: buildings, rivers or farms). The classification of the data would generally be allocated to the layer as a whole, and not to the individual features in the layer. In general, it was impossible to have sub-classes within a layer. Each layer would also be given its own *symbolology*, that is, how the data in the layer are represented graphically when portrayed on a display screen, on a printed map, etc. Symbolology includes the colours, line styles, hatching styles, symbols and fonts used, as well as the placement of labels, grid lines and other things over the geographical data being portrayed on the map. This data model also corresponded well with printing processes at the time, and it was not unusual for a layer to be referred to by its symbolology (eg: the *green layer*) rather than by the geographical data it contained. For more details on symbolology, see Section 2.9.3.

The real disadvantage of a layer-based GIS is the problem of providing topology across layers. Invariably, geometry is duplicated across layers, and whenever changes are made to the database, a process has to be run on the database to ensure that the coordinates match up across layers. Hence, in addition to the redundant storage of coordinates (which can be significant), and the less rigid topology, layer-based GISs require additional post-processing whenever changes are made.

### 2.5.3 Feature-based geographical information systems

For the next generation of GISs, it was realised that the geographical data should not be made dependent on their geometry or symbology, but that it should be the other way around. For this, the abstract concept of a *feature* was introduced (see Section 2.3.5 above), to which the feature's *attributes* would be attached: descriptions and characteristics of the feature, together with its geometry and symbology. Unsurprisingly, some GISs were hybrids between layer-based and feature-based systems, with the layer being used for a higher-level class (for example, *Roads*), and within the layer, features being assigned to more refined classes (for example, *Main Road*, *Secondary Road*, or *Track*).

The CSIR was a pioneer with the development of the feature model during the 1980s, through its GISs *KANDELAAR I*<sup>18</sup> and *KANDELAAR II*<sup>19</sup>. These systems laid the foundations for the development of the commercial GIS, *ReGIS* (which was bought by the American company, Autodesk Inc, in 1995 and has since evolved into the product, *AutoCAD Map* [GIM International 2010]), and the South African *National Exchange Standard* (*NES*), for which I was the chief architect [van Biljon 1987; Cooper 1988b; Clarke *et al* 1987; Scheepers 1989; Cooper 1989a, 1993; Cooper & Hobson 1991].

### 2.5.4 Object-oriented geographical information systems

*Object-oriented* programming was introduced primarily through Simula67 in the 1960s and Smalltalk in the 1970s, but really took off during the 1990s, with languages such as C++ and the development of graphical user interfaces. The first successful and widely-used object-oriented GIS probably was SmallWorld, released in 1989. A narrow definition of object-oriented programming does not exist, unfortunately, with the term covering variations such as *class-based* programming (where inheritance comes from classes of objects) and *prototype-based* programming (where inheritance is cloned from existing objects that serve as prototypes), and many such programming languages also catering for procedural programming [Wikimedia 2016]. Confusingly, the term *object-based* has also been used, as a limited form of object-oriented: essentially in describing a GIS, object-based means the same as feature-based, see Section 2.3.5.

Referring to the mandatory features (golden rules) in the manifesto of Atkinson *et al* [1989], and to Rigaux *et al* [2002], an object-oriented GIS then adds the following functionality, at least, to a feature-based GIS.

- **Encapsulation.**

This prevents any external code from interfering with the data or code (behaviour or methods) within an object. A feature-based GIS will generally allow access directly to all the components of a feature: attributes, geometry, symbology, operations, associations, metadata, etc. Hence, the feature is not encapsulated.

---

<sup>18</sup> A comprehensive computer-assisted cartographic system with a gazetteer, handling alphanumeric and vector data; with facilities for managing maps, aerial photographs and flight plans; and integrated with a comprehensive system for handling raster data, including the generation and draping of perspective views.

<sup>19</sup> A comprehensive GIS that was designed, but for which only parts were built

## 2. Spatial data infrastructures and geospatial data

---

- **Methods.**

These are the procedures associated with an object. As above, a feature-based GIS will generally allow any function to be performed on any feature instance.

- **Overriding, overloading and late binding.**

Late binding means that operation names (the methods) can only be resolved at run time, not at compile time, because they depend on the data types of the parameters for the operations. Clearly, this makes type checking difficult. The advantage is that it allows for overriding or overloading an object, that is, having a single name for a set of operations that can behave differently, depending on the data. It also means that one can define an operation before knowing of all the possible data types for which the operation needs to cater: as each new data type gets added, its required functionality gets added. This is implementation-specific.

However, I disagree with Atkinson *et al* [1989] that overriding, overloading and late binding are always desirable: while they can seem elegant, they can also be confusing and complex. An example might be trying to define an attribute *depth* uniquely across all feature types: the depth of a well might be the vertical distance down from ground level, a bathymetric depth might be the vertical distance below mean sea level, but the depth of a building might be the planimetric distance from its front to its back.

- The other mandatory features of Atkinson *et al* [1989] are implementation-specific and most are what one would expect of the geospatial database in the GIS:
  - **Computational completeness**, that is, the required functions can be written;
  - **Extensibility**, meaning that new, user-defined data types can be added;
  - **Persistence**, meaning that objects and their data survive to be executed again;
  - **Secondary storage management**, being ‘invisible’ performance features such as index management, data clustering and data buffering;
  - **Concurrency**, that is, simultaneous access by multiple users;
  - **Recovery**, against hardware or software failures; and
  - **Ad hoc query facility**.

## 2.6 Formal models

During 2000, the then Spatial Data Standards Commission of the International Cartographic Association<sup>20</sup> began their research on spatial data infrastructures [Cooper & Nielsen 2000]. The Commission decided to build formal models of an SDI, using the Reference Model for Open Distributed Processing (RM ODP) [ISO/IEC 10746-1 1998; ISO/IEC

---

<sup>20</sup>I have been active in the Commission since 1999 and I am the Past Chair. It is now known as the Commission on Geoinformation Infrastructures and Standards and is Chaired by Prof Serena Coetzee of the University of Pretoria.

## 2. Spatial data infrastructures and geospatial data

10746-2 1996; ISO/IEC 10746-3 1996; ISO/IEC 10746-4 1998] to provide the framework, and the Unified Modelling Language (UML) [ISO/IEC 19501 2005] to model the details. RM ODP provides five *viewpoints* for considering a system, within which any technique can be used to model the system:

- **Enterprise Viewpoint:** the purpose, scope and policies for an SDI. It describes the relationship of an SDI to its environment, its role and the policies associated. See Figure 2.1 for the Commission’s high-level model of an SDI from the Enterprise Viewpoint [Hjelmager *et al* 2008].
- **Information Viewpoint:** the semantics of information and information processing incorporated into an SDI. It could define conceptual schemas (formal descriptions of the model) and methods for defining application schemas.
- **Computation Viewpoint:** a functional decomposition of the SDI into a set of services that interact through interfaces. This captures the details of these services and interface definitions without regard to distribution.
- **Engineering Viewpoint:** the mechanisms and functions required to support distributed interaction between the services and data within a system (ie: the SDI). This is concerned primarily with the interaction between distinct services and data. Its chief concerns are: communication, computing systems, software processes, and the clustering of computational functions at physical nodes of a communications network.
- **Technology Viewpoint:** the specific technologies chosen for the implementation of an SDI [Hjelmager *et al* 2008].

The Commission modelled SDIs from RM ODP’s Enterprise and Information Viewpoints [Cooper *et al* 2003a; Hjelmager *et al* 2005, 2008], and from RM ODP’s Computation Viewpoint [Cooper *et al* 2007, 2009b, 2012c]. The Commission has not attempted to model SDIs from the Engineering and Technology Viewpoints, as they are implementation-specific. These models were developed to cater for the typical SDIs at the time, which generally were dominated by data from official data providers. The Commission then considered how its models catered for VGI, focusing first on the the general roles of stakeholders within and around an SDI (*Policy Maker, Producer, Provider, Broker, Value-added Reseller (VAR) and End User* [Hjelmager *et al* 2008]) and in the process, identified 37 special cases of these general roles [Cooper *et al* 2011c].

Other authors have found these models to provide useful frameworks for understanding SDIs, such as Makanga & Smit [2008]; Putra [2010]; Mansourian & Abdolmajidi [2011]; Béjar *et al* [2011]; Lopez-Pellicer *et al* [2011]; Oliveira & Lisboa Filho [2015]. As part of the project “*Modelling a national health spatial data infrastructure for Namibia*”, funded by the South Africa-Namibia Joint Technical Research Partnership Programme Bilateral Agreement, I have been helping colleagues at the Universities of Pretoria and Namibia use these models of an SDI to understand the Namibian Spatial Data Infrastructure (NamSDI), which is currently under development [Sinvula *et al* 2012, 2013]. Then, with colleagues at the University of Ghana, we applied the models to study the SDI stakeholders in Ghana [Owusu-Banahene *et al* 2013]. For an unrelated project, I also found these ICA models to

## 2. Spatial data infrastructures and geospatial data

be very useful as the basis for a systematic comparison of the SDIs in several countries [Cooper *et al* 2014].

As its name implies, UML brought together several modelling techniques and graphical notations that were used for object-oriented software engineering. As a result it is comprehensive but complex, with alternative ways of modelling concepts. UML has 14 different types of diagrams, but in many cases only the *class diagram* is used, to show the structure of a system: see Figure 2.1 for an example. ISO/TC 211 uses both RM ODP and UML in its standards.

### 2.7 Data quality and metadata

Perhaps the most common objections raised against VGI are uncertainty over the quality of the VGI and poor documentation of the data (eg: see Chapter 7 and Cooper *et al* [2010a]). There is a close coupling between *quality* and *metadata*, which is why they are discussed here together. Some consider quality to be part of metadata — at least, the reporting of quality. Some aspects can also be difficult to split between quality and metadata, such as *currency* and *lineage* (see Section 6.5). I do not believe that it is necessary to have a perfect allocation of concepts between *quality* and *metadata*, as long as the concepts are documented in one or the other. In general, contributions to a repository of VGI need to conform to various standards specified for the repository, particularly concerning the structure or syntax of the data: see Section 5.8.

Quality and metadata can both be *explicit* (identified and documented) or *inferred*. Indeed, it is possible that many users of geospatial data (whether professionally produced or VGI) make assumptions about the quality of, and metadata about, a data set — whether or not they have actually been documented. For example, it is likely that few read the metadata printed on a 1:50 000 map sheet.

Conceptually, the issues affecting the quality of VGI should be the same as those for professionally generated geospatial information. However, there are differences, for example, some VGI lacks adequate metadata, or the metadata is not readily available. Similarly, some folksonomies for VGI can be unreliable or reflect a narrow view of the world. The ready availability of cheap and reasonably accurate GNSS receivers means that the positional accuracy of VGI recorded using such a receiver should generally be accurate enough for most consumer-oriented purposes, such as navigation and recording points of interest. Typical errors that are likely to occur with amateur use of a GNSS receiver are transposing coordinates (quite easy to do in South Africa, because the coordinate values for latitude and longitude are similar for a large part of the country), or using the incorrect reference surface [Cooper *et al* 2012a]. However, see Section 6.4 and Figure 6.3 for examples of other problems that can occur with GNSS.

Goodchild [2008b] suggests that *accuracy* is an essential component for data integration (such as in a mashup), “because the measurement of location cannot be perfect, and hence two independently determined estimates of the location of any feature on the Earth’s surface will not agree”. In fact, it is not just knowledge of the *positional accuracy* that

## 2. Spatial data infrastructures and geospatial data

is essential for integrating data sets, but knowledge of the other dimensions of quality as well (*attribute accuracy, semantic accuracy, temporal accuracy, completeness, logical consistency* and *lineage*), as discussed in Section 6.5. Further, what is often more important is the *relative* quality of different data sets or different features (which Goodchild [2008b] terms *binary metadata*), rather than their *absolute* quality, because the data might be used in a local context. For example, if some vector data are aligned to a mis-registered (or even unregistered) image using control points in both data sets, the resulting composite might be perfectly adequate for printing out as a map to use in the field, but generally could not be combined with other data geocoded independently. Spatial dependences between features effectively mean that one cannot insert corrections into a data set that are only partial and/or independent, because of the risk of unacceptable topological errors [Goodchild 2008b].

Unfortunately, it is very easy for errors to propagate and persist, even long after they have been “removed”. An example is the non-existent *Sandy Island* in the Coral Sea: first recorded on a map in 1908, it was probably a pumice raft seen in 1876. Even though it has been repeated “undiscovered” since, it still appears in respected databases such as the General Bathymetric Chart of the Oceans (GEBCO) [Seton *et al* 2013; Achenbach 2013].

The issues concerning data quality, including the seven dimensions thereof, are discussed in detail in Section 6.5, and the details of metadata are discussed in Section 5.2.

### 2.8 Incremental updating and versioning

As an SDI delivers integrated base data sets obtained from various sources, end users use these integrated data as the spatial, temporal and other frameworks on which they build their own, value-added data sets. Ultimately, these data sets are much more important to the end user than the base data. Unfortunately, the base data sets are updated asynchronously and by different custodians, which can corrupt the reference frameworks for the value-added data. This is the complex problem of the *incremental updating and versioning* of spatial data [Cooper & Peled 2001; Cooper *et al* 2003b; Peled & Cooper 2004].

Previously, I wrestled with the problem as a Co-Chair<sup>21</sup> of the Commission of Incremental Updating and Versioning of the International Cartographic Association. Unfortunately, we found it to be a very difficult problem and while some members of the Commission made progress on narrowly-defined problems within their own domains (for example, Arnold & Wright [2005]), collectively in terms of the general problem, we got no further than defining what the issues are [Peled & Cooper 2004]. The basic problem is that the user builds their value-added data on top of the base data sets and their topology, and any changes to the base data could break the links to the value-added data, change the position of the value-added data, etc.

The value-added data could also depend on transient details in the base data. For example, Figure 6.5 shows VGI on Google Earth, purportedly of pirate boats on the beach at

<sup>21</sup>With Prof Ammatzia Peled, University of Haifa, Israel.

## 2. Spatial data infrastructures and geospatial data

---

Eyl in Somalia [“expedition” 2009]. However, the boats might then be at sea when the updated image is loaded on Google Earth and the VGI would then point to an empty beach, or even worse, to ‘innocent’ fishing boats [Cooper *et al* 2010a].

While professionals would usually capture and integrate data according to a plan, which would (hopefully) cater for some aspects of facilitating incremental updating and versioning (at least within their own data), this is often not likely to be the case with VGI. Hence, as VGI is often contributed piecemeal, the problem of linking up the updates to different bits and pieces of VGI is that much more complicated.

## 2.9 Cartography

### 2.9.1 The nature of cartography

The International Cartographic Association (ICA) has the following definitions in its mission statement:

- A **map** is a symbolised representation of geographical reality, representing selected features or characteristics, resulting from the creative effort of its author’s execution of choices, and is designed for use when spatial relationships are of primary relevance.
- **Cartography** is the discipline dealing with the art, science and technology of making and using maps.
- **Geographic Information Science (GI Science)** refers to the scientific context of spatial information processing and management, including associated technology as well as commercial, social and environmental implications. Information processing and management include data analysis and transformations, data management and information visualisation [International Cartographic Association 2003].

This definition of *map* correlates with the model of *real and virtual maps* [Moellering 2000], as shown with examples in Table 2.3. The *real map* has a permanent cartographic image; the *virtual map type I* has a transient cartographic image (a temporary map); the *virtual map type II* is permanent but cannot be viewed directly; and the *virtual map type III* is neither directly viewable nor a permanent reality, but can be transformed easily into the other states. Moellering [2000] suggests that the 16 possible transformations in this model “define *all* the cartographic and spatial data processing steps that exist in cartography, and the spatial sciences”.

Essentially, cartography is about the user interface to spatial data (whether or not in an SDI), which has many different complexities. Concerned over the limited awareness of how maps communicate that they have encountered within the SDI community, Hopfstock *et al* [2013] propose a methodology for effective map making within an SDI. Hence also, the ICA has a wide range of Commissions, dealing with issues such as art, specialist representations (eg: of mountains, of other celestial bodies, for and by children, or for and

## 2. Spatial data infrastructures and geospatial data

Table 2.3: Real and virtual maps [Moellering 2000]

		Directly viewable?	
		Yes	No
Permanent reality?	Yes	Real map eg: <i>Paper map</i>	Virtual II eg: <i>CD-ROM</i>
	No	Virtual I eg: <i>Visualization on screen</i>	Virtual III eg: <i>Spatial database</i>

by the visually impaired), specialist applications (eg: crisis management), generalization, projections and map production.

### 2.9.2 Limitations of cartographic representation

In the same way that data are not a perfect model of the real world (as discussed in Section 2.3.4), a map (in whatever form) is not a perfect representation of the data themselves. The principal problem is that the Earth is an oblate spheroid, not a plane, so the data need to be projected onto the map. Further, the cartographer needs to select what to display and how: hence, every map is always partial (reflecting the conscious and unconscious biases of the compilers) and sometimes does not show all the available data because to do so would render the map illegible. Maps can also be subjective [Panchaud *et al* 2015].

The construction of the map might also require some modifications to ensure the essential data are visible, such as by shifting apart features that would obscure one another (eg: a road and railway), aggregating some instances of a feature type, but not others (eg: aggregating buildings in a city into a new feature type, *built-up areas*, but not isolated buildings in rural areas), or omitting some instances of a feature type to make clearer other features that the cartographer considers more important. Such legitimate manipulations of the data are known as *cartographic licence*. Finally, the cartographer needs to make decisions about the *symbolology* for representing the data (see Section 2.9.3), the *annotations* (which could be in multiple languages and/or character sets) and the *map furniture* (the graphics and text that surround or overlay a map, such as titles, scale bars, North arrows, legends, coordinates, grid lines and copyright statements). Muehlenhaus [2013] has proposed a taxonomy of the rhetorical styles of persuasive cartography (or geo-communication, as he terms it), that go beyond the “neutral” cartographic licence in selecting what to represent and how to do so. The cartographer also needs to understand the needs and abilities of the map’s audience: Schmitz *et al* [2015] reports on a small study we did on the maps the CSIR has produced for use in court and in criminal investigations.

So, any map is always only just one of many possible ‘views’ of the data it portrays, which is (hopefully) documented in the metadata for the map itself.

## 2. Spatial data infrastructures and geospatial data

---

### 2.9.3 Symbology

Symbology includes aspects such as the symbols, colours, line styles, hatching patterns, shading, fonts and other devices used to represent data on a map. The following are some of the issues which have to be considered when creating symbology, as presented in Cooper [1993]:

- The context in which the map is likely to be used (eg: available illumination), the skill levels of the likely audience, and perception and ergonomic issues, such as the visual acuity of the intended audience (colour blindness and other visual impairments), the sensitivity of their feeling or hearing (for tactile or audio maps), etc.
- Combinations of colour, density, style, etc.
- Label placement, compass roses, scale bars, grid lines, legends, text and other allied information.
- Scale-dependent and -independent symbology.
- Point, line, area and solid (three-dimensional) symbols, followers, fills, hatches, etc.
- Catering for both vector- and raster-based symbols.
- Topological relationships within the symbology (such as symbols overlapping and symbols composed of a combination of other symbols).
- Capabilities and limitations of the various output devices.
- Ensuring ease of use, versatility and completeness.

### 2.10 Virtual globes and geobrowsers

Today, the term *virtual globe* is most often used to refer to a client application that provides masses of digital geospatial data in the form of a globe over the Internet, the best-known example being *Google Earth* [Google 2016a]. However, a virtual globe does not have to be available online: in 1998, Microsoft released *Microsoft Encarta Virtual Globe 1998 Edition* that allowed users to browse a 3-D model of the Earth seamlessly, that was stored on one's computer [Microsoft News Center 1997].

A *geobrowser* is a client application for accessing a complex infrastructure of software and geospatial data behind the scenes [Craglia *et al* 2008], that is, the software that allows a user to view digital geospatial data over the Web. Geobrowsers have become rather full of cartographic detail and multimedia, leading to visual clutter, information overload, steep learning curves and deteriorating system performance. This is exacerbated by the profusion of VGI [Çöltekin & Clarke 2011b]. Following Harvey [2009], particularly concerning Chapters 7 and 8 of this thesis, a *virtual globe* will be regarded here as the software-based representation of the world in the form of a globe. If the *geobrowser* presents the geospatial data as a globe, then it is also a virtual globe. Conversely, if the virtual globe is presented over the Web, then it is also a geobrowser. The terms *geobrowser*

---

## 2. Spatial data infrastructures and geospatial data

---

and *virtual globe* have been used interchangeably when referring to Google Earth [Butler 2006; Craglia *et al* 2008; Goodchild 2008c; Graham 2010].

In the context of VGI, it is important to note that the data repository is distinct from the software, that is, the virtual globe or the geobrowser through which it is viewed. This distinction is noted by Google in its terms of service for example, stating that one may only access or use the content (ie: the geospatial data) through technology (ie: a virtual globe such as Google Earth) authorized by Google [Google terms of service 2010]. Potentially, should the commercial interests so allow, a virtual globe or geobrowser could access several different data repositories, and one data repository could be accessed by several different virtual globes or geobrowsers. Thus, the same set of VGI can be viewed through any geobrowser or virtual globe. Virtual globes are a major conduit for disseminating VGI, and hence are closely coupled with VGI.

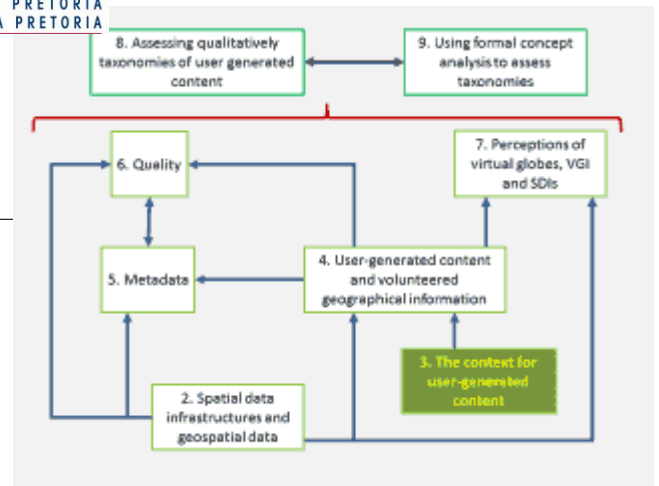
A geobrowser is the interface to a virtual globe, typically allowing users to zoom into the data, switch data layers on and off, create three-dimensional views and add their own data (user generated content), such as geospatial features (eg: roads and places of interest), tags (with text or links to Web sites) and photographs. A virtual globe can be seen as a form of an SDI (facilitating the availability of, and access to, geospatial data) as competition for an SDI, (providing vast quantities of data for free) and/or as sources of VGI (much VGI is available through Google Earth, for example).

### 2.11 Summary and looking ahead

This chapter has discussed spatial data infrastructures in South Africa and elsewhere; the terminology, types and complexities of geospatial data; classification, ontology and their encoding, including the curse of clever codes; models used in GISs; formal models of SDIs; data quality and metadata; incremental updating and versioning; cartography; and virtual globes and geobrowsers. Together, these provide the context for understanding SDIs, VGI and how VGI can contribute to an SDI; and the context for the assessment of repositories of VGI, done in Chapters 8 and 9.

Chapter 3 provides further context for the rest of this thesis and all the subsequent chapters draw on it. It will now provide details of the context that made the proliferation of user-generated content and volunteered geographical information, and the development of spatial data infrastructures possible: inter networking (which is much more than just the Internet and the World Wide Web), services, the semantic Web, social mapping, (impossibility of) controlling the Internet, open archives and access, privacy, censorship, liability, patents, copyright, curation, the digital divide and standards. All of these impact on VGI and SDIs. Together, Chapters 2 and 3 provide the setting for the subsequent chapters, but they also make important contributions as part of my research and this thesis.

\*\*\*\*



## Chapter 3

# The context for user-generated content

### 3.1 Overview of the chapter

In Chapter 2, I discussed spatial data infrastructures (SDIs), geospatial data, classification and related concepts. This chapter provides further context for the rest of this thesis and all the subsequent chapters draw on it. Specifically, this chapter examines what made the proliferation of user-generated content and volunteered geographical information and the development of spatial data infrastructures possible, and the impact of such fecundity: inter networking (which is much more than just the Internet and the World Wide Web), online services and content, the Semantic Web, social media, privacy, censorship, liability, the right to exploit content, curation, the digital divide and standards.

- Section 3.2 provides an overview of the development of *inter networking*, looking at the good, the bad and how inter-networking magnifies behaviours.
- Section 3.3 discusses the *dot.com bubble* and Section 3.4 what came *after the bust*: mashups, archiving, portals, agents or bots, Web scraping and harvesting, search engines, cloud computing, collaboration software, software as a service, syndication, mobile computing, exploiting the long tail, the Internet of things, games, the sharing economy, replacing analogue services and the deep Web.
- Section 3.5 introduces the *Semantic Web*, which has the intention of linking together content from within and across documents and repositories, such as SDIs.

### 3. The context for user-generated content

---

- Section 3.6 discusses *social media services*, of which there is a great variety. As well as recreation and entertainment, such services can hopefully also promote social justice.
- Section 3.7 introduces *social mapping*.
- Section 3.8 presents an important contribution of this thesis, namely why it is not possible for anyone to *control the Internet*. It discusses the perceived power of the Internet and the possible future, and in between presents some reasons why the Internet cannot be controlled: the sheer volume of content, the availability of technologies and data from different sources, open and secure archives, encryption, anonymity, hidden vs public transcripts, equipment, expertise, alternative forms of telecommunications, electrical power, funding, conflicting national interests, leaks, generational change and activists.
- Section 3.9 discusses *privacy, censorship and liability*, which are often used as excuses for one another. This section covers covert surveillance, covert hacking, trans-jurisdiction surveillance, overt surveillance, becoming accustomed to the surveillance society, mutual surveillance, making data already in the public domain more visible, processing available data, opting in vs opting out and assuming one has nothing to hide. Liability is often poorly understood — particularly concerning UGC.
- Section 3.10 discusses the *rights to exploit content*: patents, copyright and open access.
- Section 3.11 presents an overview of *curation*, particularly digital curation, which is a function of an SDI. I have a concern over the curation of content from Africa.
- Section 3.12 discusses the *digital divide*, which can be made worse by an enforced telecommunications monopoly or pseudo-security. This section also considers the size of the digital divide and presents some research issues, published in Cooper *et al* [2011b].
- Section 3.13 provides an introduction to *standards*, and their relationship to VGI.

The major original contribution that I have made that is presented in this chapter is:

- To information science, explaining why the Internet cannot be controlled, in spite of the best (or worst) intentions of those who should know better. Specifically, as explained in Section 3.8, there are many alternatives for communicating, finding content, hiding content, power and funding.

Additionally, the key contributions that I have made that are presented in this chapter are:

- An overview of privacy issues, see Section 3.9.1;
- Highlighting critical problems with the Protection of State Information Bill [South Africa 2013c] that do not appear in other analyses of the Bill and which draws on my invited presentation Cooper [2011a], see Section 3.9.3; and

### 3. The context for user-generated content

---

- Exposing some myths on the causes of the digital divide, see Section 3.12.

This chapter also raises some questions for further research:

1. In Section 3.5, what role does metadata play in actually enabling the linkages within linked data, and the linkages when integrating linked data?
2. In Section 3.5, can the concept of linked data be extended to metadata to create linked metadata, that is, linking items in metadata with one another?
3. In Section 3.4.12
  - (a) Does the 1% rule make the Web more radical, because the 1% that are the creators of original content have very different perspectives from the 90% that are lurkers?
  - (b) Does the need to stand out from the silent majority and defend one's position encourage more extreme positions?
  - (c) Does this create and reinforce filter bubbles, as such unsparing attitudes discourage engagement and encourage users to seek out safe harbours where the opinions and declarations match their own perspectives?
4. Drawing on Cooper *et al* [2011b], possible research issues concerning the digital divide and VGI, geovirtual environments and SDIs presented in Section 3.12 include:
  - (a) How can SDIs, geovirtual environments and other repositories of geospatial data address *information poverty* and the digital divide?
  - (b) Do virtual globes and other repositories of VGI entrench or exacerbate the digital divide?
  - (c) How can geospatial services on mobile devices help users understand their spatial context and impact on others?
  - (d) Does too much bandwidth actually result in lower-quality VGI, effectively providing quantity rather than quality?
  - (e) How should a virtual globe decide how to prioritise the data that can be displayed [Cooper *et al* 2011b]?

## 3.2 Inter-networking technologies

### 3.2.1 The good ...

Computer networks have been in existence for over 50 years, starting with military projects such as SAGE (Semi-Automatic Ground Environment), a system that connected radar stations to central control centres for tracking enemy bombers. During the 1960s, computer networks evolved from being only dedicated networks with a narrow range of tasks, such as SAGE, into general purpose networks. During 1962 at Bolt Beranek and Newman, Inc, JCR Licklider (who had worked on SAGE) developed his ideas for the

### 3. The context for user-generated content

---

*Intergalactic Computer Network*, which he subsequently presented to the Advanced Research Projects Agency (ARPA) [Licklider 1963], where he was the first head of computer research [Leiner, Barry M and Cerf, Vinton G and Clark, David D and Kahn, Robert E and Kleinrock, Leonard and Lynch, Daniel C and Postel, Jon and Roberts, Larry G and Wolff, Stephen 2003]. This led to the implementation of ARPA's ARPANET in 1969, the first production network connecting together heterogeneous computers. In turn, this led to the development of *inter-networking*, the connection of multiple independent computer networks of arbitrary design via *gateways* or *routers*.

The development of the Transmission Control Protocol/Internet Protocol (TCP/IP) in 1973 allowed other computer networks to connect to ARPANET and to each other, and the Internet was born [Leiner, Barry M and Cerf, Vinton G and Clark, David D and Kahn, Robert E and Kleinrock, Leonard and Lynch, Daniel C and Postel, Jon and Roberts, Larry G and Wolff, Stephen 2003]. On 24 October 1995, the American Federal Networking Council (FNC) came up with the following definition of "Internet", after consulting widely:

*The Federal Networking Council (FNC) agrees that the following language reflects our definition of the term "Internet".*

*"Internet" refers to the global information system that –*

- 1. is logically linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions/follow-ons;*
- 2. is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols; and*
- 3. provides, uses or makes accessible, either publicly or privately, high level services layered on the communications and related infrastructure described herein* [FNC 1995].

However, the Internet Protocol Suite (TCP/IP and related protocols) is not the only set of protocols for inter-networking, and others were used by networks such as USENET, BITNET and FidoNet. The major computer companies also had other technologies, such as IBM's Systems Network Architecture (SNA), which was launched in the early 1970s and is still widely used. ARPANET itself migrated from the Network Control Protocol (NCP) to TCP/IP on 1 January 1983 and only in 1985 did the National Science Foundation (NSF) in the USA enforce TCP/IP on NSFNET, to be followed by the other networks. Key to the success of the Internet was the use of open standards and free access to the basic documents. The developers of the Internet made use of *requests for comments (RFCs)* for their documents, such as protocol standards, comments on them and background information [Leiner, Barry M and Cerf, Vinton G and Clark, David D and Kahn, Robert E and Kleinrock, Leonard and Lynch, Daniel C and Postel, Jon and Roberts, Larry G and Wolff, Stephen 2003].

One of the myths about the Internet now is that it is the only inter-network. While one can expect lay people to be ignorant about this, it is surprising how many "experts" in the field have this delusion. See Section 3.8 for a discussion of this.

### 3. The context for user-generated content

---

The *Internet* is but one implementation of *internet* technology, such as TCP/IP. Other implementations of internet technology include the military networks in many countries, that have no connection to the Internet. The technologies, standards, documentation and devices needed to set up a network based on internet technology are readily available, and there are many who have the skills to set up their own “Internets” for whatever reason they might want to do so.

As mentioned above, *internet technology* is but one way of doing *inter-networking*, and there are still other technologies being used for inter-networking today. For example, I first made use of email through a FidoNet [FidoNet 2016] account at Rhodes University in the late 1980s. FidoNet is a point-to-point and store-and-forward email network using dial-up modems and with gateways to other networks and FidoNet was intended “to be a cooperative anarchy to provide minimal-cost public access to electronic mail” [Bush 1993]. FidoNet allowed us to send emails to users on other networks. Set up initially to support communication between bulletin board systems (BBSs), mainly through short dial-up sessions, FidoNet is still being used, particularly in Russia and the Ukraine [FidoNet 2016].

Initially, the main functions of the Internet were to enable researchers to access resources on remote computers (such as specialized software or powerful computers), exchange files (through FTP, the file transfer protocol) and exchange email. Then, mailing lists were added, followed by online games, news groups, BBSs, instant messaging and chat rooms. By the mid 1980s, the number of institutions connected to the various inter-networks and the number of users started increasing dramatically, resulting in an even bigger increase in the information available online. The Internet grew from 4 hosts in 1969, to 188 in 1979 and 159 000 in 1989, and from 837 networks in 1989 to over 134 000 by 1996 [Zakon 2006]. This growth spurred the development of indexing and searching systems, such as Archie, Gopher and the Wide Area Information Servers (WAIS), and the use of markup languages, particularly the *Standard Generalized Markup Language* (SGML), for labelling parts of a document.

At the same time, Tim Berners-Lee developed the network-based *hypertext* system, the *World Wide Web* (WWW), using the *Hypertext Transfer Protocol* (HTTP) for communicating between clients and servers. However, to take off, that needed the development and dissemination of *browsers*, such as Mosaic. SGML spawned a variety of widely-used markup languages, such as the *HyperText Markup Language* (HTML), used for Web pages, and the *Extensible Markup Language* (XML), for encoding documents so that they are both human and machine readable. The WWW grew from one Web site in 1990, to 23 500 in 1995 and over 25 million in 2000 [Zakon 2006]. By 2015, the WWW was estimated to have 47 billion Web pages [Dewey 2015a].

Through until 1989, ARPANET, NSFNET and the other major networks (such as JANET, the academic network in the UK) were closed to commercial traffic, though there were other, public, networks that carried commercial and/or private traffic from the late 1960s, such as Telenet, Tymnet, CompuServe, BITNET and Usenet. For many years the inter-networked community was small and the research and academic networks were largely self-regulating (which worked well), through the documented concept of *netiquette* (net etiquette) [Templeton 1991; Hambridge 1995], which was developed over the 1970s and

### 3. The context for user-generated content

1980s in response to network abuses and mistakes.

#### 3.2.2 And the bad

Gradually, commercial pressures increased — I remember well, unfortunately, when the small American law firm of Canter & Seigal<sup>1</sup> performed the first, massive commercial spamming of many Usenet news groups on 12 April 1994. Similar was the political spamming of news groups about the Armenian massacre by “Serdar Argıç” (an alias), also in early 1994 (which I also remember), with many of ‘his’ postings probably being produced automatically by a program scanning news groups for certain key words [Wikimedia 2016].

Malicious activities on the Internet also began in the 1980s. Although experimental and mostly harmless self-replicating programs (computer viruses and the like) were written and released during the 1970s, the first malicious and destructive computer viruses (eg: Elk Cloner, ARF-ARF, Brain and Ping-Pong<sup>2</sup>) appeared during the 1980s, though spread primarily through sharing diskettes, rather than through computer networks [Wikimedia 2016]. The first significant worm, developed by Robert Morris, was released in 1988 and led to the first successful prosecution for such as offence [US Court of Appeals for the Second Circuit 1991].

Spam and scam emails are very common, with many being labelled as 419 scams, from the relevant clause in the Nigerian Criminal Code Act in Part 6, Division 1, Chapter 38, *Obtaining Property by false pretences; Cheating*. Clause 419 covers “*any person who by any false pretence, and with intent to defraud, obtains from any other person anything capable of being stolen, or induces any other person to deliver to any person anything capable of being stolen, ... any person who by any false pretence or by means of any other fraud obtains credit for himself or any other person*” [Nigeria, Federation of nd]. As well as such criminal activities, some people create fake lives on the Web to compensate for whatever deficiencies they might have. As an experiment, Zilla van den Born showed how easy it was to pretend to be on a trip in Asia, without leaving Amsterdam [Reynolds 2014].

Really troubling is digital communication that is harmful, particularly *cyber-bullying* or harassment over the Web, and *trolling*. A survey by the Pew Research Center reporting that 40% of users had suffered harassment [Gross 2014]. These actions include verbal abuse, humiliation, discrimination, threats of physical violence, stalking and sexual harassment, such as “revenge porn”: publishing online compromising or sexually-explicit photographs of the victim, such as by a former partner. The harassment takes place primarily on social media Web sites and applications, then in the comments section of Web sites, in online gaming, by email, on discussion sites, and on dating sites and applications, and much harassment is anonymous [Gross 2014]. Besides vendettas, cyber-bullying is also used to target celebrities and activists, such as women promoting gender equality

<sup>1</sup>Unsurprisingly, Canter had previously been charged with neglect, misrepresentation, misappropriation of client funds and perjury [Wikimedia 2016].

<sup>2</sup>Combating Ping-Pong, also known as Bouncing Ball, led to the development of the CSIR’s successful anti-virus product, VPS.

### 3. The context for user-generated content

---

[Minter 2014; Cilliers 2014]. The harmful practices can include faked photographs, lies and blackmail [Addley 2014]. Cyber-bullying of children has led to truancy, poor school performance, mental health problems, self-harm and even suicide [Burrows *et al* 2012].

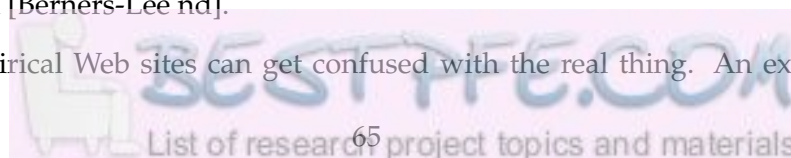
Some of those who attract hatred on the Internet might well deserve it, of course, such as those who lie blatantly to promote their online services or themselves, plagiarists, child molesters and animal abusers. All of these appear on the list of the 15 most-hated people on the Internet in 2015, of Dewey [2015b]. Unsurprisingly, this list is topped by Martin Shkreli, who increased the price of an AIDS drug from US\$ 7.50 to US\$ 750.00 per pill (and was subsequently arrested for securities fraud), followed by Walter Palmer, the killer of Cecil the Lion [Dewey 2015b].

Buckels *et al* [2014] define online trolling as “*the practice of behaving in a deceptive, destructive, or disruptive manner in a social setting on the Internet with no apparent instrumental purpose*”. The troll is similar to a practical joker and they “operate as agents of chaos on the Internet”, targeting current and divisive issues (eg: religion, politics, race and gender) to make others “appear overly emotional or foolish in some manner”. Unsurprisingly, the trolling can increase and become more vicious when someone takes the bait [Buckels *et al* 2014]. Trolling can be a form of satire (see The Onion examples below) — or just nasty. For the experienced Internet user with some gumption, trolling is easy to spot and ignore, but for naifs or those who are careless or react too quickly, it can make the Internet horrible and dangerous.

Myths have also become prevalent, sometimes starting as true stories that evolve and get warped, as happens in the game of broken telephone. Craig Shergold does exist, did have cancer and did want to get into the Guinness Book of World Records in 1989 for receiving the most greeting cards. Within a year he had received 16 million cards and in 1991 his cancerous tumour was removed successfully. However, the cards (of various forms) keep flowing in to various addresses (over 200 million, so far) and the details change, sometimes to those of real, dying children, but without their knowledge — until they get the avalanche of cards [Mikkelsen 2014].

Further myths are that Al Gore or Tim Berners-Lee invented the Internet or claim to have invented the Internet. In the case of Al Gore, he played a leading role in promoting the funding that enabled the likes of ARPANET to evolve into the publicly and commercially used Internet. However, in an interview on 8 March 1999, he stated a bit carelessly: “during my service in the United States Congress, I took the initiative in creating the Internet” [Kessler 2013]. One of the Internet pioneers, Vinton G Cerf, does believe that Gore “deserves significant credit for his early recognition of the importance of what has become the Internet” [Cerf 2000]. Fiveash [2014] found five news reports identifying Berners-Lee as the inventor of the Internet — yet these reports were celebrating the 25th anniversary of the Web! The Internet Hall of Fame inducted Berners-lee as the inventor of the World Wide Web [Internet Hall of Fame 2012], but he himself states that he did not invent the Internet and that he connected the ideas of others (namely TCP/IP, the domain name service (DNS) and hypertext) to create the Web, but the real challenge was getting others to join in [Berners-Lee nd].

Supposedly satirical Web sites can get confused with the real thing. An example was



### 3. The context for user-generated content

an article on the American satirical Web site, The Onion, declaring the leader of North Korea, Kim Jong-un to be “The Onion’s Sexiest Man Alive for 2012”: it resulted in a 55-page photo spread of Kim in the online version of the Chinese Communist Party’s official newspaper, The People’s Daily [BBC 2012b]. Then, Figure 3.1 shows the former FIFA Vice-President, Jack Warner, seemingly holding up a copy of the Onion article that fooled him, “FIFA Frantically Announces 2012 Summer World Cup In United States”<sup>3</sup> [The Onion 2015], while berating the United States of America for its fraud investigations into FIFA [Topping 2015]. Part of the problem in cases such as this is the differences in humour, culture and language across the world. Further, the satire, puns or other humour might just be trite, vacuous or otherwise not obvious.

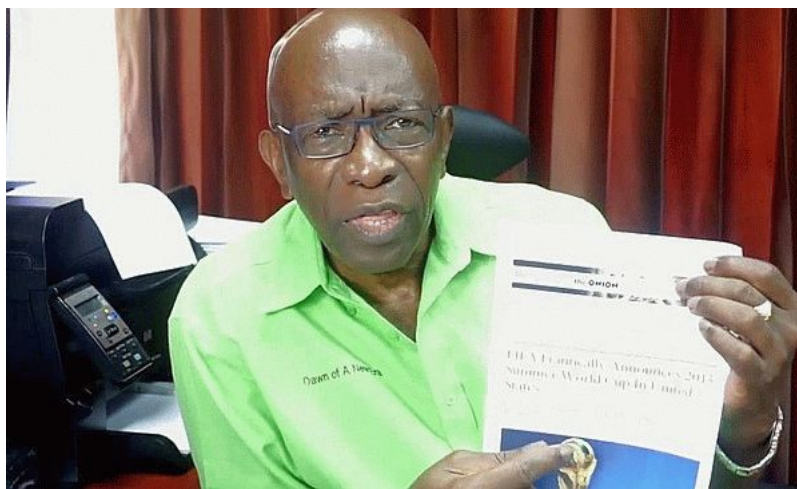


Figure 3.1: Jack Warner seemingly holding The Onion that fooled him [Topping 2015].

#### 3.2.3 The magnifier

A common mis-perception is that the Internet and the World Wide Web have changed human behaviour. They have merely exaggerated human behaviour and brought it to a wider audience. All the good and the bad of human behaviour on the Internet and World Wide Web have analogue precedents, such as chain letters, bait and switch scams, advance-fee scams, impersonation and counterfeit cheques [Wikimedia 2016]. *“The distinguishing feature of electronic communication is that it has the capacity to spread beyond the original sender and recipient, and envelop the recipient in an environment that is pervasive, insidious and distressing”* [Burrows et al 2012].

Hence, while some tactless remark in spoken conversation would often wither away (particularly without any hard evidence of it), as a tweet on Twitter it becomes immediately available world-wide, with consequences that can be catastrophic to the careless

<sup>3</sup>So obviously a spoof, as “at press time, the U.S. national team was leading defending champions Germany in the World Cup’s opening match after being awarded 12 penalties in the game’s first three minutes” [The Onion 2015].

### 3. The context for user-generated content

---

commentator. Communities that would have been impossible in the analogue world because of distance, time, costs (not just financial, but also in terms of effort, etc) and hence ignorance of one's fellow travellers, become trivial to find, join and create online, no matter how specialised or obscure they might be. Indeed, not only does the arcane become "mainstream" and even profitable to retailers through the *long tail* [Anderson 2004] (see Section 3.4.12) but *filter bubbles* [Pariser 2012] become more prevalent.

On the one hand, the Internet, Web and social media can bring geographically-separated families closer together in cyberspace and can enable parents to monitor more closely who their children have as friends and what they say to one another. On the other, it can separate, as family members indulge in social media while together, rather than talk to one another or participate in family pursuits, bombard one another with photographs of babies and parties (sometimes from narcissism) and expect grandparents to text new-born babies to bond with them [Tett 2015].

### 3.3 The dot.com bubble

Hence, the beginning of the 1990s saw the explosion of the Internet as it became readily available to companies and the lay public; provided hyper-links, graphics and pictures through the World Wide Web; provided masses of data and the indexing and searching systems necessary to find relevant information; and began to be exploited by the likes of spammers and pornographers. A key enabler was the *convergence* of computing and telecommunications, in terms of carriers (broadcast radio and television, fixed-line and mobile telephony, cable and satellite television, and computer networks), architectures, content and applications: voice over internet protocol (VOIP), games, video on demand, location-based services (LBS), multimedia, spin-offs (eg: a book spawning a film, sequels, games, comics, Web sites, advertising, clothing and toys) and user-generated content (fan fiction, parodies, etc). Convergence depends on the migration from analogue to digital<sup>4</sup>, which frees up bandwidth, facilitates integrating/transferring content and applications across platforms and allows access to multiple channels from one device: mobile telephone, game console, television, etc.

The result was the *dot.com* boom, as entrepreneurs found new ways to exploit the Internet to provide new services, such as:

- **Gopher**: a lightweight protocol for distributing and searching for documents, primarily through a text interface;
- **Netscape**: the first widely available browser;
- **Yahoo!**: an hierarchical Web directory, structured manually, established in 1994;
- **Lycos**: a search engine established in 1994, which also identified the top Web sites on the Internet;
- **WebCrawler**: for full-text Web search, established in 1994;

---

<sup>4</sup>Much delayed in South Africa: the switch over was due by 17 June 2015.

### 3. The context for user-generated content

---

- **Infoseek**: a complex system of search modifiers, established in 1994;
- **GeoCities**: a Web hosting service, established in 1995;
- **AltaVista**: a search engine, crawler and indexer with **Babel Fish**, a natural-language translation engine, established in 1995;
- **eBay**: an online auction and shopping site, established in 1995;
- **Craigslist**: a centralized network of online communities, featuring free online classified advertisements, established in 1995;
- **Salon**: an online magazine with a discussion board community, established in 1995;
- **Inktomi**: a Web-crawler search engine, used in **HotBot**, both established in 1996;
- **Amazon**: an online shopping Web site, established in 1996;
- **Google**: a search engine, established in 1996, which introduced page ranking in 1997;
- **PayPal**: an online payment and money transfer system, established in 1998;
- **Open Directory Project (ODP)**: an hierarchical ontology scheme for organizing site listings, established in 1998; and
- **Netbot**: a price comparison shopping service [Wikimedia 2016].

As can be seen, several of these pioneers still dominate their market niches (which some turned into massive markets), though they have evolved. The explosion of the Internet, World Wide Web and the offerings on them led to reckless investing in the dot.com bubble, which burst, unsurprisingly, at the end of the 1990s.

## 3.4 After the dot.com bust

In spite of the dot.com bust, the development of online services continued. The mechanisms available for communicating and sharing content have expanded, such as different forms of instant or short messaging (Twitter began in 2007), the ability to add multimedia content to messages, video telephony (Skype was released in 2003), Web conferencing, file sharing (both illegal and legal) and social networking and online communities. *Social media* is discussed in more detail in Section 3.6.

Unsurprisingly, advertising online has expanded dramatically, and is a major source of revenue for many Web sites (as it has been for traditional media). This also includes sponsored content or *click bait*, presented as pictures and links to gossip, news and other sites alongside or below the content of a Web page. One such provider, Taboola, generates revenue of US\$ 250 million annually from its click bait [Smale 2014b].

Other developments are described below.

### 3. The context for user-generated content

---

#### 3.4.1 Mashups

A *mashup* is an aggregation of existing services and/or content (data) — mashed up together — which could be purely derived or could contain original contributions. This raises issues of copyright and fair use: whether *user-derived content* or *user-copied content* [Gervais 2009], as discussed in Section 8.4.3.

#### 3.4.2 Archiving of Web pages

While phenomenal amounts of content are made available online daily, they do not just disappear into the ether. Not only do social media services retain their content (as their users presumably expect), but *archiving* of Web pages is done by a variety of organisations for various reasons.

- Commercial services archive for legal reasons, such as for proving compliance to laws regulating disclosure of material information that could affect share or other prices.
- Commercial and in-house services archive to preserve an organisation’s heritage.
- National archives do so for legal deposit, such as in the United States of America [Shiels 2010] and the United Kingdom [Sillito 2013].
- Non-profit services, such as the Internet Archive and its *Wayback Machine*, aim to provide “universal access to all knowledge” [Internet Archive 2016a,b].
- Intelligence services archive in vast quantities for reasons of national security, paranoia or whatever, as exposed by Edward Snowden in the USA and UK [Gellman *et al* 2014; Friedersdorf 2014; Farivar 2014].
- Individuals archive for reasons of curiosity or their personal research. For example, I have saved well over 500 Web pages as potential references for this thesis.

Archiving is complicated by legal issues, such as copyright, and technical issues, such as accessing the deep Web. For example, in France the Bibliothèque nationale de France (BnF) is responsible for the legal deposit of the French Internet. Such “national libraries are perceived as trusted third parties capable of creating rationally-constructed and well-documented collections, but such archives raise certain ethical and methodological questions” [Stirling *et al* 2012]. These include who can have access to what; selection and defining a meaningful corpus; privacy, such as the social media content from a partially closed circle; improper use; volatile communities, broken links, changing addresses, pages that disappear and dynamic content; the dark Web; classifying Web pages and retaining metadata, such as the popularity and reputation of a Web page. A researcher needs to cite a source and should justify its selection — but anyone who understands the Web knows how limited their knowledge is [Stirling *et al* 2012].

Further, some have the naïve assumption that content made publicly available on the Web can be expunged permanently at a whim. The European Court of Justice decided that anyone has “the right to be forgotten” can require a search engine to remove certain

### 3. The context for user-generated content

---

pages from its search results for specified terms (eg: a person's name) [Court of Justice of the European Union 2014], going against the advice of its Advocate General [Court of Justice of the European Union 2013]<sup>5</sup>. This has obviously been used by the unscrupulous to hide their activities. While such pages do not get deleted, they are removed from the search engine.

As a result, perfectly legitimate reporting by respectable organisations such as the BBC gets proscribed, contravening the public interest: for example, see Peston [2014] and Lane [2014]. Essentially, this is defaming the author of such a proscribed article by declaring their work to be illegitimate. Further, it requires the search engine's operator to make the decision over what is a legitimate removal request, and what is not, which is inappropriate [Zittrain 2014b; Scott 2014b].

A new trend is for some applications to provide ephemeral data to (hopefully) protect one's privacy, that is, content that gets deleted permanently after a specified time. They are meant to promote more authentic communication. Examples are SnapChat for photographs and Silent Circle for two-way transmissions of voice, email, video, etc. However, there is doubt that ephemerality can be enforced securely [Shein 2013].

#### 3.4.3 Portals

A *portal* is a Web site providing comprehensive and consistent access to content. A portal can be *horizontal*, that is, providing broad access to similar things from different sources, such as across an economic sector, to similar organisations or types of data; or *vertical*, that is, providing specialist access within a niche or narrow field, but often with a variety of content such as news feeds, expert opinions, publications or even e-commerce, especially business-to-business. An example of an horizontal portal is South Africa Government Online<sup>6</sup> and an example of a vertical one is for ISO/TC 211, *Geographic information/Geomatics*<sup>7</sup>.

#### 3.4.4 Agents or bots

An *agent* or a *bot* is a software tool that runs automated and semi-automated tasks over the Internet. Typically, such tasks are simple, repetitive and/or need high throughput, though agents are becoming more sophisticated. Some functions of agents are Web indexing, harvesting content or metadata, maintaining Web sites (eg: as is done for Wikipedia and OpenStreetMap) or searching for specific content (eg: police searching for illegal activities). Bots can also be used for malicious purposes, such as spamming, cracking secured Web sites or stealing content.

---

<sup>5</sup>Unfortunately, the Court issues only one judgement and no dissenting opinions, and all deliberations of the Court are secret. As I have pointed out in an email to the Court's press office, this encourages bad law by forcing all the judges to support the majority opinion and protects incompetent judges from public scrutiny.

<sup>6</sup><http://www.gov.za/>

<sup>7</sup><http://www.isotc211.org/>

### 3. The context for user-generated content

---

#### 3.4.5 Web scraping and harvesting

*Web scraping* consists of taking content from other Web sites, which is also known as *harvesting*. The collecting can be targeted and pre-arranged, such as the harvesting of metadata for an SDI from its data providers, see Section 5.2. Such Web scraping is also the actual intention of the *Statistical data and metadata exchange (SDMX)* standard [SDMX 2009], which is for collecting national statistics but is incorrectly described as a metadata standard, see Sections 6.6.4 and 5.7.

The collecting can be done using well-behaved bots (as search engines do for indexing the Web), or it can be done by simulating a human accessing Web sites. This raises issues of copyright, such as decided in the so-called “Google Defense” case concerning thumbnails of images on a pornography Web site [US Court of Appeals for the Ninth Circuit 2007]. a current case in South Africa involves Moneyweb claiming that Fin24 copied its news stories illegally [Evans 2015].

#### 3.4.6 Search engines

A search engine obviously has to do some form of Web scraping to locate the content first, before it can be classified and indexed to provide the rapid search responses that users expect. There are different types of *searching engines*, which use diverse paradigms such as semantics, ontologies or faceted classification, with the most successful being Google. A recent technique is dynamic computational analysis, as used by WolframAlpha [WolframAlpha 2016]. To return the results as quickly as they do, the results from search engines are not always accurate (particularly the results count) and there is much of the Web they cannot access [Alexander 2012].

A whole industry has now been built up around *search engine optimization (SEO)*, which is aimed at increasing the likelihood of Web pages been accessed by users searching for content. The key is to get the target page as high up as possible in the results returned by the search engine. Google even provides a “Search Engine Optimization Starter Guide” [Google Search Engine Optimization Starter Guide 2010], covering issues such as unique and accurate page titles, using the ‘description’ meta tag, the structure of the site, sensible URLs, site navigation, site maps, content and services, images and mobile-friendly sites. It appears that there is much policing done by SEO practitioners of the false and misleading activities of others, such as addresses for service area businesses (SABs)<sup>8</sup> that are really outside of its service area, and might even be in a place where they are not licensed to operate. There seems to be a particular problem over this with locksmiths in the USA [Austin 2014].

SEO does not only affect commercial decisions, such as which hotel or restaurant to use, but also other consumer attitudes, behaviours and choices, such as voting. In their study of the power and robustness of what they call the *search engine manipulation effect (SEME)*, Epstein & Robertson [2015] found “that (i) biased search rankings can shift the voting

---

<sup>8</sup>A SAB is one that travels to deliver a service, such as a locksmith or plumber, rather than one that provides its service at its premises.

---

### 3. The context for user-generated content

---

preferences of undecided voters by 20% or more, (ii) the shift can be much higher in some demographic groups, and (iii) such rankings can be masked so that people show no awareness of the manipulation”. Unsurprisingly, SEO can also reinforce or break the filter bubble [Bakshy *et al* 2015; Pariser 2015].

#### 3.4.7 Cloud computing

*Cloud computing* consists of accessing remotely through a network, particularly the Internet, computing resources such as hardware (eg: high-performance computing such as clusters or grids, or storage, be it primary storage or for backups), software (especially specialist or expensive software) or even remote printers (eg: large-format, high-volume, high-quality or 3D printers). Typically, users do not know where the resources are that they access from the cloud, and these could even change dynamically, depending on charge rates, usage, etc. To some extent, cloud computing is a return to the era of centralised (but now out-sourced) mainframe computing and thin clients (dumb terminals), but where resources are rented as and when needed and where resources are (hopefully) maintained and kept up to date.

Potential problems with cloud computing include privacy, surveillance by cloud hosts and national spying agencies, security, sustainability, and with the varying data protection and other legislative requirements in different jurisdictions, compliance and other legal problems. It is likely that many users of cloud services are unaware of these issues, or perhaps even unaware that they are actually using cloud services. This is perhaps illustrated by the leaking online in August 2014 of nude photographs of over 100 stars, allegedly obtained from a cloud storage service used for backing up the contents of mobile telephones [Butterly 2014]. Ordinary citizens are reluctant to use available security features because of their complexity, do not understand the defaults of their devices and applications they have downloaded and do not understand where copies of their images and data are [James 2014].

Further, for most users, cloud computing is always likely to be more expensive, slower and less reliable than buying one’s own computers and storage [Doctorow 2009]. Cloud services also need to be managed carefully, to ensure that one is not wasting resources [Boulton 2015]. Cloud services are also vulnerable to take-downs (as happened to Megaupload [BBC 2012f]), when users could lose access to their data and services in the cloud, or even lose them completely and permanently. Cloud services can also fail, as happened to Amazon’s EC2 in April 2011, which disrupted some well-known services for several days [BBC 2011a].

#### 3.4.8 Collaboration software

*Collaboration software* enables teams (whether together or remote) to work together, such as wikis (a Website allowing users to modify its content through a Web browser), shared calendars, conferencing (Web, video and audio), workflow management, shared folders and shared applications (particularly word processing and spreadsheets).

### 3. The context for user-generated content

---

#### 3.4.9 Software as a service

*Software as a service (SaaS)* is software and data hosted in the cloud and made available on demand, particularly for enterprise business applications. Unfortunately, SaaS can often consume a lot of bandwidth. This might be because such services get developed in environments with excellent bandwidth — which might actually be a precondition for developing a SaaS! I recall how a decade or so ago, South Africa’s abysmal connectivity stymied pioneers trying to establish such Web services.

#### 3.4.10 Syndication

*Syndication* is the sharing of Web feeds such as breaking news (whether licensed or free) across Web sites and to end users through services such as the Really Simple Syndication (RSS) and its extension for geospatial data, GeoRSS. Syndication is an easy way to expose one’s content widely and an easy way for Web sites to provide current and dynamic content passively.

#### 3.4.11 Mobile computing

*Mobile computing* means using devices such as portable computers, personal digital assistants (PDAs), laptop computers, mobile telephones, tablets and wearable computers. Initially, mobile computing was a matter of being able to take one’s computing power, software and data to other locations, such as into the field. Unsurprisingly, mobile devices are now far smaller, far cheaper, far more powerful and far more capable (with GNSS receivers and a variety of sensors), and able to connect to the Internet from most anywhere, through cellular telephony, short-range wireless data services (eg: Wi-Fi and Bluetooth) or even satellite services. Now, mobile computing is also about location awareness, being mobile while using the device and a variety of user interfaces, such as touch screens and voice input.

Of the nearly 5-billion mobile phone subscribers worldwide, over 2-billion are using smartphones [Bremmen 2015]. A smartphone is “a mobile phone that performs many of the functions of a computer, typically having a touchscreen interface, Internet access, and an operating system capable of running downloaded apps.” [Oxford 2016]. Many users now access the Internet and the Web only through mobile devices, so it is common for Web sites to have mobile-friendly versions, catering for the smaller screen on the mobile device, etc.

#### 3.4.12 Exploiting the long tail

*Exploiting the long tail* means using digital media and the Internet to overcome the traditional constraints of geography and scale to aggregate niche markets, so that their specialist tastes can be served. Retailers such as Ecast, Amazon and Rhapsody make lots of money from sales of material too arcane for traditional retailers to keep on their

### 3. The context for user-generated content

---

shelves [Anderson 2004]. This also permits mass customization (especially now with the looming proliferation of three-dimensional printers) and encourages distributed, open or user-driven innovation. At the end of the tail, though, an arcane group can still be so large that its members think that they are interacting with a diverse audience, and hence create a filter bubble.

Analogous to the long tail, the bulk of contributions to UGC repositories such as Wikipedia is made by a small proportion of the contributors. Similarly, most of contributions to a VGI repository such as OpenStreetMap are made by few participants. This is known as the 1% rule of thumb, whereby about 1% of the Internet community create all the original content, about 9% edit or otherwise contribute some content, and the remaining 90% 'only' view or consume content, that is, are lurkers [Wikimedia 2016]. Obviously, this is a simplification and the actual ratios will vary across different fora and communities. However, it does raise several questions.

- Does the 1% rule make the Web more radical, because the 1% that are the creators of original content have very different perspectives from the 90% that are lurkers?
- Does the need to stand out from the silent majority and defend one's position encourage more extreme positions?
- Does this create and reinforce filter bubbles, as such unsparing attitudes discourage engagement and encourage users to seek out safe harbours where the opinions and declarations match their own perspectives?

#### 3.4.13 The Internet of Things

The *Internet of Things* (IoT) is based on the proliferation of objects (not just computers, tablets or smartphones) that are connected and identified uniquely, all combining to provide smarter services. [Rifkin 2014] considers it to be the convergence of the communications, energy and logistics Internets, connecting everything and everyone. He foresees a struggle between global companies (telecommunications, Internet, energy and electricity) trying to enclose the IoT and monopolise the flows, keeping prices high, and the 'prosumers' creating lateral networks (the collaborative commons) to share at near-zero marginal costs.

Many smartphones, for example, have microphones, cameras and barometers; motion, orientation and proximity sensors; and location determination (through GNSS, Wi-Fi, GSM, compasses, etc). There is already a variety of other sensors that can be plugged into smartphones, such as for monitoring pollutants. While the Internet of Things can have many positive benefits, there can be negative consequences, such as invasion of privacy, interfering with other rights through constant monitoring, trust, security, governance and standards [O'Reilly & Battelle 2009; Coetzee & Eksteen 2011, 2012]. Technology can often let one down, so everything private about one could become public (or owned by law enforcement or marketeers) and/or all one's systems (eg: an insulin pump) could be hacked or used for ransom [Sullivan 2013]. Ramirez [2015], the FTC Chair, identified three to the privacy of consumers that could undermine their trust: ubiquitous data col-

### 3. The context for user-generated content

---

lection, unexpected data uses with adverse consequences and security risks. She feels these should be addressed by “security by design”, data minimization, transparency and informing consumers of unexpected data uses and giving them choices.

There is a concern that the benefits of the Internet of Things has been exaggerated. For example, Dlodlo *et al* [2013] propose that the Internet of Things can be used to reduce crime in South Africa, but some of their proposed solutions have nothing to do with the Internet of Things, such as linking databases for fraud detection, crime mapping and analysis, and money laundering. More importantly, the real problem is not the lack of technology, but people issues, such as the lack of skills and capacity, and corruption.

#### 3.4.14 Games

The playing of computer or video games online dates back to the 1970s, such as on the PLATO<sup>9</sup> (Programmed Logic for Automatic Teaching Operations) computer-based education system, thought they were obviously limited by the bandwidth then available, and hence often text-based. By the 1990s, the spread of the Internet, cheaper computers and faster bandwidth made online gaming more feasible and more attractive, such as with better graphics. From around 2000, video game consoles could connect to the Internet. Online gaming then grew rapidly, with persistent, massively multi-player online games (MMOGs); the ability of players to enter and exit an ongoing game at any time; new genres of games such as social games; abstract games; games without specific objectives; virtual goods that can be traded for real money; professional game champions; access from mobile devices and cross-platform gaming. Problems with online games include cyberbullying, violence and xenophobia (see [Rakitienskaia *et al* 2011; Rakitienskaia 2015] for more details) [Wikimedia 2016].

#### 3.4.15 The sharing economy

As discussed in Section 4.6, the Web has facilitated *crowd sourcing*, that is, soliciting services from the crowd. In a similar vein, the Web has facilitated the *sharing economy*, whereby individuals share capital items and other resources (for money, barter or even paying forward<sup>10</sup>), using the Web to find one another. The result is that the owner of the resource effectively amortizes their investment more rapidly, while those requiring only occasional use of such a resource do not need to make the capital investment. Nominally, it also provides access to a supposedly lucrative industry, such as tourism, and potentially cheaper holidays for tourists [Monks 2014].

Unsurprisingly, this has created an opportunity for brokers to facilitate the exchange, such as Uber<sup>11</sup> and Lyft for transport (as competition for, or in collaboration with, taxi drivers) and AirBnB for accommodation. Peer-to-peer rentals are estimated to amount

<sup>9</sup>To which I was exposed while at Rhodes University in the early 1980s.

<sup>10</sup>That is, the contributor donates the value with the expectation that the recipient will do the same for someone else in the future.

<sup>11</sup>Which has already arranged over 1 billion rides [AFP Agency Staff 2016].

### 3. The context for user-generated content

---

to US\$ 26 billion annually [Monks 2014]. This is all promoted as being the sharing economy, though Stallman [2014a] states that it is really a *piecework subcontractor economy* — as some forms of crowd-sourcing are. Currently, Uber drivers have embarked on a class-action suit against Uber in California, over whether they are employees or independent contractors, and to recover tips that Uber withholds from them [Levine 2015].

What has never been clear to me is whether or not these brokered services are actually covered by the insurance taken out by the provider of the service: normal vehicle insurance does not even cover the use of one's car for business trips, even short ones within a town, never mind covering the running a taxi service. Then "*Uber's clever policy of not being directly responsible for anything that goes wrong extends to harassment by drivers, and its practice of identifying passengers enables drivers to find out who the passenger is*" [Stallman 2014a]. Further, it appears that these brokers can be quite aggressive, morally bankrupt and misogynist. Lacy [2014], a female journalist, documents a threatened illegal smear campaign against her by Uber, because of her reporting on Uber.

The sharing economy is disrupting traditional industries, such as taxis, and sometimes ignoring valid legislation and regulations concerning health, safety and employment conditions [van Zyl 2015a; Fischer-Baum & Bialik 2015; AFP Agency Staff 2015a; Finkelstein 2015]. Without such controls (created over decades as people died in fires, traffic collisions or collapsed buildings, from food poisoning, unsafe equipment, etc), both the suppliers and consumers can be endangered or inconvenienced, such as being raped by a driver [Barry & Raj 2014], being stalked [Schneier:2014 2014], or being detained in a drug-lab raid of one's rented accommodation [BBC 2016]. Newman [2015] warns that the powerful brokers in the sharing economy can also pose a threat to local democracy, using their wealth to fight politicians through attack advertisements, employing large teams of lobbyists and exploiting their big data collections on their 'partners', customers and transactions to target voters. There is also a concern that the sharing economy will reduce tax collection, as the service providers under-declare their incomes because such services are not their main source of income [AFP Agency Staff 2015c]. In South Africa, Uber is probably already reducing the fleet renewals of car rental firms [Peters 2015]. However, WesBank has entered into an agreement with Uber to rent cars to the "driver-partners" [van Zyl 2015b].

On the other hand, Finkelstein [2015] suggests that there are opportunities for applications and services allowing the 'partners' to share information seamlessly and hence empower themselves against the broker, and to exploit "the inherent power imbalance that has characterised the first wave of the sharing economy ... disrupting the disrupters?"

#### 3.4.16 Replacing analogue services

Gradually, online digital services have been replacing analogue services, such as teletext services: the first one launched (on 24 September 1974), BBC Ceefax, was switched off on 23 October 2012 [Hand 2012]. The pioneering French online service with dedicated terminals, Minitel (providing the telephone directory, banking, shopping, stock prices, weather, government services, etc), was switched off on 30 June 2012, after exactly 30

### 3. The context for user-generated content

---

years of service [Schofield 2012]. Indeed, with all the rapid changes, some consider the Internet and the World Wide Web to be in *perpetual beta* [O'Reilly 2005]! While the likes of search engines and portals provided focused access into the masses of data, documents, video clips and other content on the Internet, the content was essentially unstructured and multiplying rapidly. Further, much is not accessible in the *Deep Web*:

#### 3.4.17 Deep Web

“Searching on the Internet today can be compared to dragging a net across the surface of the ocean. While a great deal may be caught in the net, there is still a wealth of information that is deep, and therefore, missed. The reason is simple: Most of the Web’s information is buried far down on dynamically generated sites, and standard search engines never find it” [Bergman 2001].

Of course, the deep, invisible or hidden Web consists not only of content generated dynamically, but also private content with controlled access (eg: password protected), particularly within an organisation’s internal networks, or unlinked and hence unreachable content. There is also much content inaccessible to search engines by convention, such as through the robots exclusion standard (robots.txt files). There are also discarded or unused resources on the Internet (such as unused IP addresses) which can be used for transient services and content [Beckett 2009].

### 3.5 The Semantic Web

“And so, it is striking to consider — almost shocking, in fact — what the world might be like when our software turns to the Web just as frequently and casually as we do. Today, of course, we can see the faint, future glimmers of such a world” [Swartz 2013]. The intention of the *Semantic Web* is to move from documents, often linked together through hyperlinks, to interconnected data that computers can consume and process. The World Wide Web Consortium (W3C) has been developing a technology stack for this of *linked data*, *vocabularies*, *queries* and *inference* [W3C 2016; Bizer *et al* 2009; Swartz & Hendler 2001].

- *Linked data*, particularly *linked open data* (LOD).  
Data are considered to be linked when they are published on the Web, machine-readable and explicitly connected, such as through chains of *Subject*  $\Leftrightarrow$  *Predicate*  $\Leftrightarrow$  *Object* triples. For this, W3C has developed standards such as the *Uniform Resource Identifier* (URI), to identify a Web resource uniquely (preferably using HTTP), and the *Resource Description Framework* (RDF), to describe or link things through subject-predicate-object triples, each of which is normally identified by a URI. RDF is a framework that can be implemented in various concrete syntaxes, such as *RDF/XML*, *RDFa* (*Resource Description Framework in attributes*), which has triples embedded in HTML or XML, or *Turtle* (*Terse RDF Triple Language*), a minimal syntax that is more human readable. *URI aliases* are different URIs referring to the same entity and when de-referenced, are the mechanism for linking disparate databases together in this *Web of Data*.

### 3. The context for user-generated content

*Linked open data* are then linked data available for reuse freely under an open licence (such as Creative Commons CC-BY<sup>12</sup>), in open, machine-readable formats that are openly available and that support RDF standards and are linked to the data of others, to provide context [Berners-Lee 2009]. Clearly, linked open data are more useful than hidden or restricted data, as they are accessible to all to be used as they wish.

- *Vocabularies.*

These are collections of classes and properties to organise and enrich data, such as through *ontologies* (see Section 2.4.4), and other knowledge organisation systems. Widely-used vocabularies implemented in RDF include *FOAF* (*Friend of a Friend*), for describing people, their activities and relationships; *DOAP* (*Description of a Project*), for describing software development projects; *vCard*, for electronic business cards; and the metadata standard for describing resources, *Dublin Core* [ISO 15836 2009]. The latter is discussed in Chapter 5, particularly in Section 5.8.8. Vocabularies can be complex because in terms of AAA: “anyone can say anything about any thing; someone will say something about every thing in every conceivable (linguistically) way” [Dunsire 2013]. Unfortunately, the interpretation of a concept can change over time or between socio-cultural contexts, a process known as *semantic drift*, which can compromise ontologies [Kontopoulos 2015].

- *Queries.*

The Semantic Web can be viewed as a global database which can be queried with an RDF-specific query language such as *SPARQL* (*SPARQL Protocol and RDF Query Language*), to match triple patterns against RDF triples. Wrappers are also available for the likes of relational databases, Web applications or Web pages, to translate SPARQL queries into other query languages (eg: SQL for relational databases) to query resources that are not encoded in RDF.

- *Inference.*

The power of the Semantic Web is being able to reason over data through rules (defined in a vocabulary), with the intention of discovering new relationships, or using the available logic in the linked data to check proofs.

According to Cox *et al* [2010], the linking of data is not merely a technical matter, but is rather a “philosophy of usage”, that is, the appropriate linkages should not just be made mechanically, but really depend on insight into the content. Linkages can also be made through widely-used naming schemes or classifications, such as ISBN and ISSN for publications, or ISCO for occupations (see Section 2.4.5). RDF links can be generated automatically, such as through ontology matching or similarity metrics [Bizer *et al* 2009]. Of course, this is dependent on the quality of the data: the example of the database of one large South African organisation with the name of the town *Witbank* recorded over 200 different ways in one field [Cooper 2007], shows how complicated this could be!

Clearly, metadata should also be published for the linked data, to enable users to assess their provenance or quality, see Chapters 5 and 6. Bizer *et al* [2009] recommend Dublin Core [ISO 15836 2009], which has been implemented in RDF; or the likes of the Open

<sup>12</sup>Attribution: allowing commercial and non-commercial distribution, remixing, tweaking and building upon, as long as credit is given.

### 3. The context for user-generated content

---

Provenance Model, which focuses on the *lineage* or history of the data (see Section 6.5), that is, agents, processes and artefacts and their relationships; or the Vocabulary Of Interlinked Datasets (VoID), which expands on Dublin Core (adding access and structural metadata and descriptions of links between data sets) to specify metadata for RDF data sets.

This raises two questions for further research:

1. What role does metadata play in actually *enabling* the linkages within linked data, and the linkages when integrating linked data?
2. Can the concept of linked data be extended to metadata to create *linked metadata*, that is, linking items in metadata with one another?

Geospatial data inherently have unspecified relationships within them that can be inferred, particularly through topology. Further, many relationships are encoded explicitly in geospatial data, through compound features, classification and associations, see Section 2.3.5. Linked data can facilitate mash-ups and they are also complementary to an SDI, allowing for navigation, sophisticated querying and reasoning [Cox *et al* 2010]. Linked data can also be used to improve maintenance processes for geospatial data sets that are inter-related but independently managed, by relating the different sources in an integrated manner [ISO/TC 211 Ad hoc group on linked data 2012], as is required for an SDI. Harvey *et al* [2014] suggest that LOD enables the third generation of SDIs.

However, it is both time consuming and challenging to model the world in triples, resulting in a knowledge base that is mostly incomplete. Any reasoning based on this will hence be limited. Nevertheless, the Semantic Web is the start for dealing with incomplete knowledge, as anyone can add to it [Hobel & Frank 2014]. Swartz [2013] was concerned that too much effort was being put into developing a massive suite of standards, rather than just building something that works, as was done with the Internet and the Web.

## 3.6 Social media services

### 3.6.1 Variety

The use of the Internet for social purposes began with email and then dial-up bulletin board systems (BBSs), for distributing news and exchanging messages and files. A well known example of a BBS that evolved into a current social media service is the *Whole Earth 'Lectronic Link*, normally shortened to *The WELL*, which began in 1985. It was then one of the first dial-up Internet service providers (ISPs) and was a key catalyst for the formation of the digital rights advocacy group, the Electronic Frontier Foundation (EFF).

The Internet and the World Wide Web has encouraged user-generated content, whereby ordinary people contribute data and opinions that get used by others. So, one could read a newspaper, or the Web site of a professional news source (eg: the BBC), or one of the many blogs out there in the 'blogosphere'. Technorati has indexed over 100 million blog

### 3. The context for user-generated content

records since 2002 [Technorati, Inc 2016], though many have ceased to exist. The distinction between a professional news Web site and a blog is blurring, as blogs become more professional and as news Web sites hosts blogs for their journalists, so many have probably used a blog without realising it. A result is that social media can be used for information warfare or to spread false rumours [Jansen van Vuuren *et al* 2012]. “A challenge news organizations face when it comes to rumor-reporting in particular is the fact that rumors tend to be much more shareable — and much more clickable — than corrections” [Garber 2014]. Generally, news items debunking rumours get very little attention, which removes the incentive for writing them [Garber 2014; Nyhan 2014]. However, journalists do have a duty to verify the news on social media [Harber 2015*b*].

As well as email, BBSs and blogs, other social media services include instant messaging (point-to-point or multicast online, real-time, synchronous text communication or chats), micro-blogs (very short blogs, eg: Twitter limits tweets to 140 characters, as with an SMS), podcasts (recorded audio or video blogs), video sharing, status updates, bookmarking, online communities with shared interests, online dating and virtual worlds. Social media services are also used in work environments, such as participating in a conference [Coetzee *et al* 2009]. It has become common to mark key aspects of content with hashtags, that is the # sign followed by a label signifying the topic, sometimes in camel case<sup>13</sup>. Hashtags are a form of folksonomy, see Section 2.4.3. They facilitate searching for, and subscribing to, content on a specific topic, such as a celebrity or event; clustering content; and making a topic ‘trend’, that is, become more prominent on search engines or social media. Hashtags are also used to integrate across media platforms, such as for an advertising campaign or for interactive commenting by viewers of a television programme. Hashtags are also used for activism, such as #BringBackOurGirls (for schoolgirls kidnapped in Nigeria) or #BlackLivesMatter (protesting against the killings of African Americans by police in the USA). They might prompt action by the relevant authorities — or might oversimplify complex issues or just be a form of *slacktivism*, see Section 3.6.2 [Louw-Vaudran 2015].

Dahlgren [2012] suggests that the social media has also changed the roles of *public intellectuals*, providing them with greater access to the public and new types of media (video, mash-ups, etc). On the other hand, there is perhaps a trend from big questions to banalities, there are defensive cultural challenges from those with unquestioning ideological convictions, and the normal discursive terrain of the humanities (that is, reflection on values and politics) is being encroached on by scientists and technologists<sup>14</sup>.

As could be expected, the distinction between different types of social media services has become blurred, with instant messaging services including video, for example, and aggregated services such as Facebook. One formerly successful social media service that originated in South Africa was MXit McLaren [2009], but it faced too much competition and failed to adapt to smartphones [Alfreds 2015*b*]. Combining the names of the services Flickr, MySpace, Facebook, Youtube and Twitter, one gets *FlickMyFaceYouTwit* [Ambrose 2009]!

<sup>13</sup>That is, with the initial letters in upper case, the rest of the text in lower case and no spaces between the words.

<sup>14</sup>As, perhaps, I am doing in this thesis!

### 3. The context for user-generated content

---

There are also now *Geo-social* networking services, for exploiting or complementing location-based services, or where locations are added to content, either explicitly (eg: a geocoded photograph or a place name) or implicitly (eg: by reverse geocoding an IP address). Such services can be used for arranging events (such as parties, flash mobs or demonstrations, by identifying those in the social network in the geographical proximity of the event), for collective situational awareness during disasters (eg: locating survivors) or for collaborative mapping, but they can also be mined for detecting trends, such as disease epidemics, or for pushing ‘relevant’ advertisements. For example, for more than a week after I left Bali in May 2012, Facebook kept on presenting me with advertisements in Indonesian!

Unsurprisingly, there is much scepticism over social media services, particularly concerning privacy, time wasting, hate speech and other inappropriate behaviour — frequently, celebrities get into trouble of their tweets, for example. Further, social media services can often require one to make explicit statements about one’s relationships and their statuses, which could be compromising and which invariably cannot reflect the subtleties, complexities and time-variance inherent in relationships and their multi-faceted nature. Every visible action sends a signal [Ceglowski 2011] — and inaction also sends a signal.

There are also criminal, offensive and anti-social activities in virtual worlds and other social media, such as harassment, impersonation, identity theft, distribution of illegal material (eg: child pornography), fraud, discrimination and, specific to virtual worlds, gold farming (acquiring large quantities of in-game currencies through intensive game playing or using bots). These activities can obviously spill over into ‘real’ life, such as the Chinese workers exploited to do gold farming [Rakitienskaia *et al* 2011; Rakitienskaia 2015]. Some suggest that terrorists love the likes of Twitter because of its ability spread their messages [Altman 2014].

Social media are used widely by “media-savvy” celebrities to enhance their brands (both their personal brand and their endorsed products), maintain their high profile and maybe even provide more than just banalities for their fans. Social media can also make ordinary people famous rapidly: see Osborne [2016] for examples of Instagram stars and Pearson *et al* [2015] for examples of YouTube stars. However, it is easy for media icons to lose their touch and show their age, as social media evolves [Robinson 2015].

On the other hand, users can use the social media to expose the celebrities (eg: disseminate unflattering photographs before they get air-brushed for publication), engendering sexist and other vitriol [Gorton & Garde-Hansen 2012]. They are also used by activists, such as “Baba Jukwa” in Zimbabwe, who appears to be a mole within ZANU-PF making leaks on Facebook, including warning of assassination attempts on the former Minister, Edward Chindori-Chininga, who was then killed in a car crash [SAPA 2013]. Other activists target the social media accounts of terrorists, their agents and fund raisers, aiming to get such accounts taken down, with Twitter removing about 2000 a week around March 2015 [Gladstone 2015].

The media can be manipulated for commercial or other purposes, as was done by the companies that made the *Harlem Shake* go viral and made money for themselves [Ashton 2013b]. It is easy to establish fake identities and create fake popularity or credibility [Ashton 2013a]. Finally, the social media services do exploit the data provided willingly

---

### 3. The context for user-generated content

---

by their users: as the aphorism goes, if you are not paying for the product, you are the product!

#### 3.6.2 Social justice

Ideally, social media services promote *social justice* and *solidarity as a form of community*, though Bassett [2013] suggests that they do not: free beer is not freedom. She proposes that one can undermine the social media monopolies through:

- *Silence* (which can be an active protest);
- Making one's texts **less** amenable to encoding through *linguistic delirium* or *glossolalia*<sup>15</sup> (by simultaneously offering meaning and refusing it: "what seemed like language slides away and is revealed as a chimera", maybe with the "potential to trick those contemporary mechanisms through which our words are captured" [Bassett 2013]);
- *Metaphors* (with their "capacity to make meanings out of impossible combinations" [Bassett 2013]);
- Or even *lying*.

Such disrupting of the language ("Hidden transcripts" vs "public transcripts" [Lugg 2013]) can also be used to resist other attempts to control the Internet (political, ideological, etc), as discussed below in Section 3.8.

A local example of social media driving social justice is the current Water Shortage South Africa campaign, to distribute water to parts of South Africa suffering under the drought: over 18 000 people had participated by mid-January 2016 [Evans 2016; Water Shortage South Africa 2016]. However, there is a concern that social justice and activism could be replaced by *slacktivism*: "actions performed via the Internet in support of a political or social cause but regarded as requiring little time or involvement, e.g. signing an online petition or joining a campaign group on a social media website" [Oxford 2016]. In democracies, slacktivism is perhaps more about boosting the ego of the participant rather than helping the cause. Because it is so quick and easy (eg: a single click on a 'like' button), it is likely that participants could forget which causes they supposedly support. Dun [2016] has set up an online petition to "Stop people on the Internet setting up frankly barking online petitions" and as one Zoe Kemp responded in supporting the petition: "I like signing something, whilst doing absolutely nothing. Now I feel warm inside".

Avaaz is an online pressure group primarily known for arranging online petitions and email campaigns, but occasionally also pickets and the like. Avaaz claim they prevented the BSkyB merger in the UK by delaying it until the News International scandal killed it in 2011. The likes of Avaaz also enable those with mobility problems to remain politically involved [Kingsley 2011]. On being diagnosed with diabetes, Busse [2015] was able to get Mattel to make diabetic accessories for the American Girl dolls with 4335 supporters — but it did take two years.

---

<sup>15</sup>Speaking in tongues.

### 3. The context for user-generated content

---

Further, slacktivism-type activities can really be defiant acts under authoritarian or repressive regimes, such as leading up to the Arab Spring. It has also been a significant factor in various protests in Turkey (eg: saving the trees at Gezi Park), because of the censorship of mainstream media [Kizilkaya 2013]. This is perhaps borne out by clamp-downs on social media, such as in Turkey [BBC 2014c].

The digital divide (see Section 3.12) can also affect social justice, with better-off communities more likely to use VGI to protect and enhance their local environment [Foster & Dunham 2014].

## 3.7 Social mapping

In common with many neologisms about the Internet, there is not a hard definition for *social mapping*. There are two ways to interpret the term *social mapping*.

- **As a form of VGI or UGC.**

That is, it could be more personal than general VGI repositories such as OpenStreetMap or Google Earth. Social mapping then includes the provision of mapping tools that can be customized themselves, as well as allowing the user to create easily maps of their data for consumption by a select group, or for posting on the Web. This form of social mapping would include community-based mapping and its variants, such as *public participation geographical information systems (PPGIS)* and *asset-based community development (ABCD)*, as discussed in Section 4.5.2.

- **As a topological map.**

That is, it could be some sort of network in either in geographical space or some abstract space, of the connections within and between social networks or virtual communities.

*Social mapping* could be a combination of both, of course, as a geospatial networking service facilitating social dynamics such as a flash mob, as discussed above in Section 3.6. This is particularly useful where the participants in the social network use mobile devices predominantly to interact with the social network, as the social mapping can then be done in real time as a location-based service. *Activist mapping*, such as applies to the origins of Ushahidi [2016], would be a form of social mapping.

## 3.8 Controlling the Internet

### 3.8.1 Perceived power of the Internet

*“It increasingly seems that the world will no longer have a single superpower, or group of superpowers, that brings order to international politics. Instead, it will have a variety of powers — including nations, multinational corporations, ideological movements, global crime and terror groups, and human rights organizations —*

### 3. The context for user-generated content

*jockeying with each other, mostly unsuccessfully, to achieve their goals. International politics is transforming from a system anchored in predictable, and relatively constant, principles to a system that is, if not inherently unknowable, far more erratic, unsettled, and devoid of behavioral regularities. In terms of geopolitics, we have moved from an age of order to an age of entropy” [Schweller 2014].*

Parts of this system are the World Wide Web and the Internet: both are exploited and feared to varying extents by all these different powers listed by Schweller. As discussed in Section 3.2, the *Internet* is merely one implementation of *internet technology*, which, in turn, is but one of many different ways of implementing *inter-networking*. Similarly, the World Wide Web is just one way of delivering vast quantities of content and services over networks. Further, as discussed below, the Internet is designed to route around disruptions.

Nevertheless, there are many (supposedly intelligent and educated) people who do not understand this and also conflate the World Wide Web with the Internet. They think that they can control **all** inter-networking activities and information flows by controlling the Internet: to censor political dissent, prevent blasphemy, hide corruption, hide illegal activities, restrict access to content, impose discriminatory pricing or tax online services. As noted by President Toomas Hendrik Ilves of Estonia<sup>16</sup>, “*We must choose between two paths — either we can change the nature of the internet by acceding to a Westphalian<sup>17</sup> regulatory structure of internet governance, or we can change the world*” [Ilves 2013].

Examples of such attempts to control the Internet include Ethiopia’s ban on using VOIP services such as Skype [Moskvitch 2012]; Wyoming attempting to ban some forms of citizen science and, for that matter, conventional tourism [Pidot 2015; Kurtz 2015] (see Section 4.4.3); China imprisoning bloggers, isolating the entire Xinjiang autonomous territory from the Internet for ten months, terminating the mobile telephone services of those allegedly using a virtual private network (VPN), and now using road blocks to check mobile phones for the likes of Skype [O’Brien 2016]; Saudi Arabia flogging and jailing bloggers [BBC 2013*d*, 2015*b*]; and the UK’s Investigatory Powers Bill or Snoopers’ Charter, requiring everyone’s browsing history to be maintained for a year for scrutiny without a warrant and seemingly requiring encryption to be weakened, which has even been criticised by the government’s own Information Commissioner [Griffin 2016]. Unsurprisingly, the Home Office considers it “vexatious” to ask for the browsing history of the Home Secretary [Stone 2015]. In December 2012, a set of International Telecommunications Regulations (ITRs) tabled at a meeting of the International Telecommunications Union (ITU) suggested that the ITU has the power to regulate the Internet [Mullin 2013*b*]. The World Press Index for 2015 lists several countries that proscribe threatening their “territorial integrity” or “national security”, particularly over the Internet, such as Russia, Morocco (particularly concerning Western Sahara), Iran, Egypt, Somalia, Ethiopia, Thailand, Burma and Indonesia. The Index also has concerns over democracies such as the USA, UK, France, Australia and Japan [Reporters Without Borders 2015].

Some private companies would also like to control the Internet, such as the mobile tele-

<sup>16</sup>Ranked first for Internet freedom for three years running [Ilves 2013].

<sup>17</sup>*Cuius regio, eius religio* (whose realm, his religion), the principle of non-interference in the internal affairs of countries, which was the basis of the 1648 Peace of Westphalia [Ilves 2013].

### 3. *The context for user-generated content*

---

phony operators in South Africa wanting to control value-add services such as WhatsApp because of their threat to their revenue [van Zyl 2016]. There are many individuals who are unhappy with the information flow and dissemination due to the Internet and Web. As the former US President, Jimmy Carter stated: “Religion, and tradition, are powerful and sensitive areas to challenge” [Carter 2009]. In Bangladesh, for example, several bloggers have been murdered in the street [Sanchez 2015; BBC 2015a]. Content providers also try to control access geographically (known as geo-blocking), to facilitate discriminatory pricing, etc.

Essentially, those wanting to control information and its distribution face problems such as the following.

#### 3.8.2 Sheer volume of content

Apparently, the National Security Agency (NSA) in the USA “intercepts and stores nearly two billion separate e-mails, phone calls, and other communications every day”, which makes the system too complex to determine whether or not it actually works [Schweller 2014], and volume stored increases daily. Other countries also have such surveillance systems, in collaboration with, or in competition with the NSA — or both. Should the system become overwhelmed by all the evidence being collected automatically, it could lose its perceived authority or effectiveness. This is through mistaking omniscience for omnipotence or intelligence, as Engelhardt [2013] suggests the NSA is doing, stating that it appears that “the more you know about the secret lives of others, the less powerful you turn out to be”! Rather than wisdom, the sheer volume creates information entropy — and information becomes noise as it “is routinely distorted, buried in noise, or otherwise impossible to interpret” [Schweller 2014].

The result might well be that such intelligence agencies create their own filter bubbles, because of the sheer volume of data they harvest, rather than in spite of the volume. Much of the content (facts, opinions, allegations, imagery, comments, conversations, folksonomies, etc) they deal with will be contradictory, so their selection, rating and analysis will invariably be biased by their preconceived notions and the desire to “simply want to believe something that feels right” [Schweller 2014].

Unfortunately, it is easy to forget the basics when one is embedded within such a massive and powerful system. This has had tragic consequences far too often, including within in the United States of America. Apparently, the American intelligence agencies ignored the warnings of their Russian counterparts about the Tsarnaev brothers and ignored the obviously suspicious activities of the brothers, resulting in the bombing of the Boston Marathon on 15 April 2013 [Investors.com 2013]. In another case, Aaron Alexis was still able to get clearance for access to the Navy Yard in Washington DC, where he shot and killed several people on 16 September 2013, even though he had arrest records for firearm offences, naval misconduct citations and known mental health problems [Leonnig *et al* 2013]. For several months before the terror attack on the Taj Mahal Palace Hotel in Mumbai in November 2008, the American, British and Indian intelligence agencies were tracking the main ring leader and other participants — even identifying the hotel as a likely

### 3. The context for user-generated content

---

target before July 2008 [Glanz *et al* 2014].

It is important to realise that being able to conduct surveillance over the Internet or being able to use the Internet to interfere with the rights of others or being able to conduct information warfare over the Internet are all quite different from being able to **control** the Internet! The genie is out of the bottle and cannot be replaced! The Internet was designed to be robust (distributed, with data sent in small packets) and self-healing if any node broke [Ananthaswamy 2011]. As the Internet pioneer John Gilmore puts it, “*The Net interprets censorship as damage and routes around it*” [Elmer-Dewitt 1993].

#### 3.8.3 Technologies

There are many alternatives to internet technology, such as UUCP, X.25 and FidoNet, and it is quite feasible to develop and use new ones that are even more robust to attempted censorship. FidoNet, for example, was intended to be a “cooperative anarchy to provide minimal-cost public access to electronic mail”, facilitated by having every node being self-sufficient, needing no support from other nodes to operate, and being able communicate with any other node as it chose, “without the aid or consent of technical or political groups at any level” [Bush 1993]. As a consequence of the mass surveillance revealed by Edward Snowden, some feel that the breakup of the Internet has already started in countries such as Brazil, India and Germany — and even a return to the typewriter for sensitive documents [Taylor *et al* 2013]. Further, two encrypted communication services that closed as a result, Lavabit and Silent Circle, have founded with others the Darkmail Alliance, to make available a service that is even more secure from eavesdropping [Hern 2013].

#### 3.8.4 Data

Alternative data sources are readily available to use to question or challenge the official line of repressive governments and exploitative corporations, such as imagery from the satellites of a variety of countries and private companies (eg: used recently to expose Sudan’s harbouring of the Lord’s Resistance Army [Ronan *et al* 2013]) or VGI such as geocoded photographs and videos taken with mobile-phone cameras. ProPublica uses the term ‘space journalism’ for its use of satellite imagery, such as to highlight rising sea levels [Albeanu 2014]. NGOs, citizen journalists and open source intelligence (OSINT) experts such as Nuba Reports and Bellingcat not only gather data and imagery, but also provide analysis and assessment of available data [Taub 2014; Higgins 2015; Seitz 2015]: see also Sections 4.3.4.1 and 4.5.2.1.

Due to insurance fraud and corrupt traffic police, for example, there is a proliferation of active dashboard cameras in Russia, which provide much user-generated content. Besides being a check on police and others, such cameras can provide interesting imagery, such as the spectacular footage of the meteor strike in the Urals in February 2013 [BBC 2013c]. Even a country as secretive as North Korea could be mapped in detail quickly by Google, when they decided to do so through crowd sourcing [BBC 2013b]. Further, even

### 3. The context for user-generated content

---

combining publicly available official data in new ways can be used to hold governments to account [McLaren 2012].

#### 3.8.5 Repositories and open archives

Even within the Internet itself, there are mechanisms for bypassing the attempts to control the Internet — even those by well-resourced countries. For example, the Web site, *Wikileaks*, uses cryptographic technologies and databases mirrored around the world (ie: within different legal jurisdictions) to provide anonymity for the likes of internal dissidents, whistle-blowers, journalists and bloggers with an outlet that it claims is “an uncensorable version of Wikipedia for untraceable mass document leaking and analysis” [Wikileaks 2016]. Wikileaks is designed to be beyond the control of repressive regimes and to remain accessible, even when it is (supposedly) blocked.

“Wikileaks was founded by Chinese dissidents, journalists, mathematicians and startup company technologists”, several of whom were involved in the Tianimen Square protests in 1989. Wikileaks aims for maximum political impact and claims to have received over 1.2 million documents in its first year of operation [Wikileaks 2016]. They claim to be able to bypass the attempts of the Chinese government to block access to Wikileaks. Clearly, a facility such as Wikileaks could be used for nefarious as well as positive purposes. Their ethical position is quite interesting, because they claim (with justification) that the existing media channels are being used to disseminate propaganda.

There are other open archives, of course, some of which pre-date Wikileaks. Examples are:

- Wikileaks spin-offs with narrow focuses, such as Brussels Leaks, Balkan Leaks, In-doleaks and RuLeaks [Wikimedia 2016];
- **Cryptome**: established in 1996 as a digital library for information on freedom of speech, cryptography, spying and surveillance, but also for banned documents [Cryptome 2016]; and
- **The Pirate Bay**: established in 2001 by an anti-copyright organisation for peer-to-peer sharing [Wikimedia 2016], it and similar sites could share politically-sensitive and other suppressed material.

The private leaking of documents or information (ie: to only a select few people or organisations) is not considered to be effective in combating corruption because it merely empowers the select few who receive the information to exploit their exclusivity. However, the public leaking of documents and information (such as through Wikileaks) makes them available to anyone, who can then interrogate, critique and challenge the veracity of what has allegedly been leaked. Of course, this privilege is open also to those whose alleged malfeasance has been exposed by the leak.

### 3.8.6 Encryption, anonymous access and anonymous communication

Effectively, most electronic communication is encrypted by being compressed, so it is meaningless to try to ban encryption. Explicit encryption is also widespread, such as for protecting financial transactions, and the tools are readily available, including as open source software. Encryption is not perfect, of course, with the National Security Agency (NSA) having installed backdoors in several widely-used and supposedly secure encryption systems [Ball *et al* 2013a; Fairweather 2013]. As Thompson [1983] showed a long time ago in his *Turing Lecture*, it is possible to introduce back doors and the like into systems without leaving a trace of them in the source code.

Encryption techniques can also hide with whom someone is communicating [Chaum 1981], with the key being eliminating the metadata trail of the transactions [Zetter 2014]. Further, the fact that encrypted information is actually being transmitted can be hidden (known as *plausible deniability*), such as through *steganography*, which is hiding secret messages within larger ones, especially those with high redundancy such as image and audio files. Image steganography techniques include using the least significant bits and exploiting the two-step compression of JPEG [Morkel *et al* 2005]. There are even techniques for denying responsibility for encrypting messages [Rivest 1998].

A virtual private network (VPN) is a mechanism for extending access to a private network from over a public network, by creating a secure, encrypted ‘tunnel’ (virtual point-to-point connection) through the public network. It is used primarily to allow the staff of an organisation to access securely its internal networks, content and services from remote sites or while roaming. A VPN can also be used to improve the robustness and security of a wireless connection. However, a VPN can allow one to circumvent censorship and geo-blocking, such as by providing one with a temporary IP address acceptable to the service being accessed [O’Neill 2012; Keall 2014].

Services such as The Onion Router (TOR) and software such as Freenet provide, to varying extents, anonymous and “untraceable” access to the Internet, not just for those in fear of a repressive regime, but also to access illegal content, such as child pornography or drugs in the *Dark, Invisible* or *Deep Web*, on sites such as Silk Road [Clarke *et al* 2001; Beckett 2009; BBC 2011d]. TOR has been and is being funded primarily by the United States Government (through the State Department and the Department of Defense) to help journalists, activists and campaigners to secure the privacy of their communications into and out of countries with repressive regimes. It is so secure, that the NSA claim they are not able to break it directly, but need to resort to exploiting vulnerabilities in browsers and other software [Ball *et al* 2013b; Schneier 2013]. Further, there is a claim that some agents in security agencies are leaking details of bugs in TOR to the team responsible for TOR [Kelion 2014].

Data can also be hidden on storage devices by using the gaps between the intended storage locations on discs, tapes, etc — which also makes it more difficult to delete data completely from such media.

### 3. The context for user-generated content

---

#### 3.8.7 “Hidden transcripts” vs “public transcripts”

The intended meaning of content can be alluded to through ambiguity: puns, spoofs, satire, pastiches, ditties, jokes and other humour. These are then the *hidden transcripts* (which can be used as *weapons of the weak*), the discourse beyond the direct observation of the power-holders, that are embedded in the *public transcript* from the subordinates. Ironically, the hidden transcripts can become icons when the authorities attempt to suppress them, such as the homophonic pun, *Grass Mud Horse*, in Chinese video spoofs [Lugg 2013].

#### 3.8.8 Equipment

Wired and wireless routers, modems, cabling, computers, other equipment and the software required to set up networks are readily available, affordable and compact, that is, relatively easy to smuggle into countries where their sale is restricted.

#### 3.8.9 Expertise

Thousands of people across the world have the technical expertise to build and run computer networks and the communication and social-media applications to run on them. The required documentation and training materials are readily available for those wanting to teach themselves.

#### 3.8.10 Telecommunications

Already, there are about 6 billion mobile telephone subscriptions in the world and 45% of the world’s population has access to 3G services [ITU 2012]. In many countries, the mobile networks are encrypted (eg: GSM) and run by private companies, and hence often beyond the control of most governments. However, there are already other communications media such as satellite telephones and ham packet radio. Further, there are initiatives such as Google’s Project Loon to provide connectivity through a network of balloons in the stratosphere [Google 2013], and Facebook’s plan to do something similar with solar-powered drones and low-earth orbit and geosynchronous satellites [Wakefield 2014]. Google and Facebook are not doing this to be altruistic, but to increase their markets and customer bases. Then “when all communication fails, you can depend on amateur radio”, as happened when an emergency occurred on Gough Island and nothing else worked: the 836-word message took 105 minutes to transfer and verify, jumping between frequencies and conditions changed [van de Groenendaal 2014].

In South Africa, the Broadband for All project had established a wireless backbone and mesh clusters to connect over 200 schools and offices and nearly 100 000 learners by 2012, creating entrepreneurial opportunities for village operators [Matthee 2012]. However, informed community co-design and local socio-political issues need to be taken into account for the design, roll-out and maintenance of such systems [Rey-Moreno *et al* 2013].

---

### 3. The context for user-generated content

---

The ‘white spaces’ in the frequency spectrum being released by the move from analogue to digital television will provide wide-area coverage at greater ranges for serving rural areas [Pejovic *et al* 2014]. In addition, 3G coverage by the mobile telephony companies in South Africa is growing rapidly [McLeod 2014].

More importantly, though, it is still possible to use mobile phones to communicate off network and anonymously, such as by building peer-to-peer networks (*mesh networks*) using Bluetooth or Wi-Fi [Leyne 2010; BBC 2011*b,d*; Simonite 2013], or even audio broadcasting of data through bursts of *digital birdsong* [BBC 2012*c*]. Apple added mesh networking to its iOS 7 operating system in 2013 [Woollaston 2014].

During the “Umbrella Revolution” in Hong Kong that began in September 2014, the activists used “fast wireless broadband, multimedia smartphones, aerial drones and mobile video projectors”, as well as mesh messaging over Bluetooth and Wi-Fi, to coordinate and provide frequent updates on social media and live streaming video [Yang 2014; Hume & Park 2014]. Information also gets shared on an *ad hoc* peer-to-peer basis, such as the micro movies popular with Chinese commuters [Sebag-Montefiore 2013]. Presumably, messages could also be broadcast using the techniques of computer worms and viruses.

#### 3.8.11 Electrical power

While electronic communication depends on electricity, it is not dependent on the power grid, because of the proliferation of small-scale, cheap, independent (off the grid) electricity generation equipment, such as solar cells and wind turbines. Further, there have been dramatic improvements in the cost, size and capacity of batteries [Hulac 2015]. Hence, information distribution cannot even be stopped by cutting off the power!

#### 3.8.12 Funding

While bartering can obviously be used to bypass authorities and their control of transactions, the development of crypto-currencies such as *Bitcoin* and *Litecoin* [Goldberg 2012; Salmon 2013; Steadman 2013] means that even money is no longer controlled by governments. Chen [2013] suggests, though, that “Bitcoin is built on a weird mix of the most old-fashioned kind of speculative greed, bolstered by a contemporary utopian cyberlibertarian ideology”. Nevertheless, crypto-currencies are a significant development, in the wake of the likes of Paypal and credit card companies refusing to honour payments to Wikileaks and other organisations. As of October 2014, there were already 285 Bitcoin ATMs around the world [Lee 2014*e*]. Indeed, to some extent PayPal now accepts Bitcoin [Harrison 2014].

Further, drawing on the successful example of the *Freigeld* of Wörgl in Austria in 1932-3<sup>18</sup>, there are local currencies to build resilience into local economies (eg: in Bristol and Totnes in the United Kingdom) [Steadman 2013; Swain 2013]. Local currencies have been stimulated by the 2008 global financial crisis and frustration over the perceived growth

---

<sup>18</sup>So successful, that it was banned by the Austrian central bank.

### 3. The context for user-generated content

---

of the wealth gap. Essentially, a local currency needs a good reason to exist, a core team of volunteers (eg: to deal with regulations), support by local businesses and residents, assets backing the currency, keeping it simple, having a design competition within the community, raising funding and then promoting the currency continually once it has been launched [Kermeliotis 2014]. There is also a variety of alternative currencies, such as frequent-flyer miles, airtime (cell phone time), loyalty cards with points, gift cards, in-game currencies and food stamps, which get traded (legally or illegally) and hence are readily used for making purchases [Swain 2013; Stein 2013]. These can now be carried on mobile devices in mobile wallets and integrated with conventional and other alternative currencies. They already provide over US\$ 160 million in purchasing power [Bonchek & Cornfield 2013].

Even though there are problems with Bitcoin, such as possible fraud concerning MtGox, then the largest Bitcoin exchange [AFP Agency Staff 2015b], or it becoming uneconomical to mine Bitcoins without cheap electricity and fast ASIC<sup>19</sup> processors [Garratt & Hayes 2015; Kaminska 2015], alternative currency mechanisms will evolve and improve. “People’s definition of money is starting to change and fragment” [Kemp-Robertson 2014], not just because of the disruptive technologies that make alternative currencies both possible and accessible, but also because there is “a general decline in the levels of trust being placed in traditional institutions” [Kemp-Robertson 2014].

#### 3.8.13 Conflicting national interests

For most countries, if not all, there are other countries motivated to expose or contradict their actions, and hence provide succour to agitators and the like. For example, both China [Mandiant 2013] and the United States of America [Greenwald 2013] have large cyber espionage programmes. Hence, those countries with the resources are likely to prevent any of their rivals from controlling the Internet.

#### 3.8.14 Leaks

If the likes of Manning and Snowden could access such ‘compromising’ secrets in such volumes and make them public, how long before them did the real spies from China or Russia (or even from North Korea, Iran, Cuba or Al Qaida) get access to those secrets, and more? As Friedersdorf [2014] stated after Snowden released to The Washington Post a large cache of personal communications taken from the NSA.

*“The same logic applies to Keith Alexander, James Clapper, Michael Hayden, Stewart Baker, Edward Lucas, John Schindler, and every other anti-Snowden NSA defender. So long as they insist that Snowden is a narcissistic criminal and possible traitor, they have no choice but to admit that the NSA collected and stored intimate photos, emails, and chats belonging to totally innocent Americans and safeguarded them so poorly that a ne’er-do-well could copy them onto thumb drives.*

---

<sup>19</sup>Application-specific integrated circuit, customized to execute specific operations.

---

### 3. The context for user-generated content

---

*They have no choice but to admit that the NSA was so bad at judging who could be trusted with this sensitive data that a possible traitor could take it all to China and Russia” [Friedersdorf 2014].*

#### 3.8.15 Generational change

The young, technically-minded people that grew up with the Internet and are shaping it, who are the likes of Edward Snowden, Bradley (now Chelsea) Manning, Aaron Swartz and Jeremy Hammond, are the ones needed by governments to be intelligence analysts and systems administrators — and are the ones who will run the intelligence agencies in a generation [Assange 2013b]. He suggests that “by trying to crush these young whistle-blowers with espionage charges, the US government is taking on a generation, and that is a battle it is going to lose”. Penny [2013] suggests that the “wildly disproportionate sentencing of young digital activists” is meant to be a deterrent, but is likely to backfire, because such hackers and whistle-blowers have little respect for the moral authority of governments and the ways they operate; they came of age “just as the financial crash of 2008 swept away the socioeconomic justification for Anglo-American imperialism”. Such moral authority has not been helped by the Anglo-American interventions in Iraq and Afghanistan [Holloway 2009]. Further, the hacker echelon is now a mix of curious teenagers, government agents, law enforcement, criminals, activists and freelancers, all with different motivations [Pagliery 2015].

Presumably, the same also applies in other countries such as China and Russia: if not right now, then rather soon.

#### 3.8.16 Activists

In *A Theory of Human Motivation*, Maslow [1943] identified five *prepotencies* in a hierarchy of needs: physiological, safety, love, esteem and self-actualization. Because of the generational change, education, ready access to like-minded people across the world (such as through social media) and literature, the greater ability to strive for the highest of Maslow’s *prepotencies* (*esteem* and *self-actualization*, as manifested in the open-source software movement, for example) and other factors, there are now many activists who have the technical skills, resources (as outlined above), time and motivations to challenge governments, major corporations, belief systems (religions, political ideologies, etc) and other groups. The activists operate on their own, in collectives (such as *Anonymous*, so loosely-associated it has internal dissent and no strictly-defined philosophy [Wikimedia 2016]) or to support communities or special-interest groups (whether marginalized or not) with legal challenges, disseminating information (or propaganda), soliciting help, or whatever.

These online activists mirror, extend, initiate, precipitate and/or counter similar types of social activism on the ground, such as the protests at the World Trade Organization (WTO) Ministerial Conference in Seattle, Washington, USA in 1999 (also known as the *Battle of Seattle*); the *Occupy Wall Street* movement in 2011 (with its slogan, *We are the 99%*);

### 3. The context for user-generated content

---

and *Restore the Fourth* [Russia Today 2013]. The Seattle protests also served to launch the Independent Media Center (Indymedia), a participatory network using open publishing to allow anyone to report on political and social issues [Wikimedia 2016].

“This, ultimately, may prove to be our strongest protection against the rise of the surveillance state. The same tools that strengthen it strengthen those who protest against it. Privacy is not the only illusion in the new age of data; government secrecy is too. Big Brother might be watching, but he is also being watched” [Von Drehle 2013]. “But before we change the world, we need to change the way we think” [Brand 2013]. Foucault [1982] suggests that “the role of philosophy is also to keep watch over the excessive powers of political rationality, which is a rather high expectation”.

#### 3.8.17 The future

Of course, this does not mean that the future will be wonderful. “We cannot ignore that the authoritarians want to encroach on the territory of the free. They want to force their authoritarianism on us. It’s our task not to let them do it” [Ilves 2013]. Private companies, political parties and wealthy individuals can also use their relative wealth against their opponents in legal action, known colloquially as a *strategic lawsuit against public participation (SLAPP)*, that aims at intimidating community groups, NGOs and activists and overwhelming their (often meagre) resources by burdening them with civil suits (often spurious and complex) to censor and silence them. Further, the uncontrollable Internet is being exploited by criminals and terrorists [Jakes & Goldman 2013; Lee 2014a]. See also Section 3.9 below, for a discussion of privacy issues.

However, freeing communications from national or international control is not merely just a civil rights, anarchic or privacy goal. The New Zealand Red Cross has developed the *Succinct Data* platform, which uses a Wi-Fi mesh between smartphones with data store and forward, to provide robust communications in disaster areas where the cellular networks are not working. Succinct Data provides real-time tracking of personnel, formatted data without transcription errors and transparent integration with the cellular network when it is available, and is easy to take across borders (unlike radios and satellite terminals) [Lloyd 2012]. As it is, there is often rampant misreporting by the media concerning disasters or emergencies (even creating unnecessary panic), such as news anchors and reporters not listening to one another, contradicting each another, exaggerating the problem, focusing on the wrong issues, using dubious sources and being susceptible to fakes [Cooper 2008b; Goldman 2013].

Perhaps as a result as their lack of control, authorities are adopting draconian measures against activists and the like, such as the persecution of the esteemed Internet pioneer Aaron Swartz (he helped create RSS and Reddit) over mass downloading of academic articles from JSTOR — which drove him to commit suicide in January 2013 [Yglesias 2013; Stamos 2013; Swartz *et al* 2013; Hellman 2013b,a]. Governments also forget just how mobile in the Internet world are people, businesses and data, and just how quickly businesses thrive or whither. Hence, the spying that the NSA has been doing using the major American Internet companies, as was exposed by Edward Snowden in June 2013,

### 3. The context for user-generated content

---

might cost those companies business as their clients move to ‘safer’ companies beyond the NSA’s control [Livingstone 2013; Masnick 2013; AFP, SAPA-DPA 2013; Sprigman & Granick 2013] — particularly as the NSA might have been conducting commercial espionage on German and other foreign companies [Moody 2013; Müller-Maguhn *et al* 2014]. The cost to US companies could be as much as US\$ 35 billion by 2016 [Taylor *et al* 2013]. The scandal might also have considerably weakened the authority of the current American dominance of the governance over the Internet [Nothias 2013; Fairweather 2013].

“It is getting to the point where the mark of international distinction and service to humanity is no longer the Nobel Peace Prize, but an espionage indictment from the US Department of Justice” [Assange 2013b].

## 3.9 Privacy, censorship and liability

There are key ethical issues concerning the Internet and its content, particularly invasion of privacy (or surveillance), censorship and liability, and they are often used as excuses for one another. For example, content could be denied or restricted (ie: censored) to “protect” privacy or because of “concern” over liability. Further, claims over the ownership of content, as discussed in Section 3.10, are also used to censor content. *“Moreover, any such agreement must also take into account the possibility that such rights, if given without adequate restraints and safeguards, may be used either by individuals or by nations to block or obstruct the communication, development, and use of literary and artistic products — a danger all too acute during a cold war period. The rights of the creator, producer, or owner must be balanced against the legitimate demands and needs of others, even if in foreign countries, to enjoy these products at a reasonable, non-prohibitive cost”* [Kramer 1954].

On the other hand, privacy could be compromised over “concerns” over liability, such as when a company monitors the emails of its staff. The issues are not necessarily well understood either, such as when “poor” data (eg: low resolution remotely-sensed imagery) is considered to be censored data, simply because it covers in inadequate detail, an area that is of interest to a conspiracy theorist, or the like. However, as noted by Ilves [2013]<sup>20</sup>, “privacy and security do not have to contradict each other; indeed, secure on-line interactions, enabled by a secure online identity, is a precondition for full internet freedom”.

### 3.9.1 Privacy

*“The right to life has come to mean the right to enjoy life, — the right to be let alone; the right to liberty secures the exercise of extensive civil privileges; and the term ‘property’ has grown to comprise every form of possession — intangible, as well as tangible”*, stated Warren & Brandeis [1890] in their seminal paper, *The right to privacy*. Their primary concern was over making private details public: “each crop of unseemly gossip, thus harvested, becomes

---

<sup>20</sup>President of Estonia.

### 3. The context for user-generated content

---

the seed of more, and, in direct proportion to its circulation, results in the lowering of social standards and of morality” [Warren & Brandeis 1890]. Subsequently, Brandeis [1928] took this further in his equally seminal dissenting judgement concerning wire-tapping in *Olmstead v United States*, stating that “it is also immaterial that the intrusion was in aid of law enforcement”.

Compromising the privacy of others can be accidental, such as the by-product of a legitimate research or VGI-gathering project. For example, while mapping a slum can help show what infrastructure is needed and the possible sources of diseases there, it could endanger some marginalised people, such as by encouraging authorities to crack down on the settlement when they realise how large it is. It is essential to include the community in such projects [Mohdin 2014]. Similarly, care must be taken when collecting oral histories, etc [MacDowell 2012].

Von Drehle [2013] suggests that “*privacy is mostly an illusion. A useful illusion, no question about it, one that allows us to live without being paralyzed by self-consciousness. The illusion of privacy gives us room to be fully human, sharing intimacies and risking mistakes. But all the while, the line between private and public space is as porous as tissue paper*”. There are several dimensions to the shattering of this illusion on the Internet, of the invasion of privacy, as discussed below.

#### 3.9.1.1 Covert surveillance

Covert surveillance is possibly what many consider surveillance to be: monitoring behaviour and communications surreptitiously, for detecting, investigating and monitoring threats (criminal, terrorist, social unrest, political, etc), influencing and controlling society, and, hopefully, protecting citizens. It appears that surveillance is becoming dominated by electronic techniques, as discussed above in Section 3.8.

Brandeis [1928] was rather prescient: “*Subtler and more far-reaching means of invading privacy have become available to the government. Discovery and invention have made it possible for the government, by means far more effective than stretching upon the rack, to obtain disclosure in court of what is whispered in the closet. . . . The progress of science in furnishing the government with means of espionage is not likely to stop with wire tapping. Ways may some day be developed by which the government, without removing papers from secret drawers, can reproduce them in court, and by which it will be enabled to expose to a jury the most intimate occurrences of the home. Advances in the psychic and related sciences may bring means of exploring unexpressed beliefs, thoughts and emotions*”.

For example, “brain fingerprinting” is claimed to detect the presence or absence of information in someone’s brain (eg: knowing what the murder weapon looks like), using electroencephalography (EEG) [Farwell *et al* 2013], though Meijer *et al* [2013] have concerns over the studies. The International Neuroethics Society considers that neuroethics includes *brain privacy*, as “we have more sophisticated imaging devices that can in a crude way begin to give clues to observers about what you’re thinking or feeling” [International Neuroethics Society 2016].

---

### 3. The context for user-generated content

However, as discussed in Section 3.8.2, it is easy to become so enamoured with such sophisticated and expensive technology that the basics get forgotten, with tragic consequences.

#### 3.9.1.2 Covert hacking

Covert surveillance also enables the removing and/or adding of evidence (ie: covert hacking) to incriminate innocent people, such as on one's private computer, or even in electronic data transmissions [Cowie 2013].

#### 3.9.1.3 Trans-jurisdiction surveillance

One feature of the designed-in robustness of a packet-switching network such as the Internet, is that one cannot guarantee the routing that individual data packets will take. Hence, even though one might have a high-speed, high-bandwidth Internet connection directly between two countries (or even between two places in one country), parts or all of the connection might be routed through other countries — and these third countries might capture and/or study the data traffic as it passes through their territory [Holputch 2013]. Such trans-jurisdiction surveillance might be accidental, of course, though it should be expected by those conducting the surveillance. For example, Internet traffic to and from the United Nations headquarters in New York is presumably routed through the United States of America and hence likely to be recorded by the National Security Agency there. An immediate result has been a rift between USA and what are usually its allies, such as Brazil [Borger 2013] and Germany [NSA-Überwachung 2013]. Even worse, it now appears that it is possible to misdirect Internet traffic deliberately and surreptitiously, particularly across national boundaries, to inspect and/or modify the data being transmitted [Cowie 2013].

Remote sensing from satellites in particular, but also from aeroplanes, facilitates surveillance of other countries. Indeed, one of the first major uses of the LANDSAT series of satellites was to monitor the wheat crop in the Soviet Union [Erickson 1984; Shurkin 2012].

#### 3.9.1.4 Overt surveillance

Not all surveillance is covert, with overt forms including those that are visible and well identified (such as CCTV surveillance cameras in public spaces, or the tedious disclaimers one gets of the call being recorded when trying to get a useful response out of a call centre) or to which one agrees explicitly (such as the small print in the terms and conditions for using a Web site). However, it is possible that in some jurisdictions, many such supposed agreements are unenforceable, being excessively long (even longer than Shakespeare's longest play, *Hamlet*!) or changed arbitrarily and without notice [Hudson 2013]. Hence, Brunon-Ernst [2015] suggests that informed consent is actually a fallacy. Some VGI might

### 3. The context for user-generated content

---

also be considered to be overt surveillance, such as private webcams in fixed and known locations.

Further, in practice it is easy to forget that one's actions are being observed, even when one has given explicit consent, as we found in projects tracking cell phones [Cooper *et al* 2009a, 2010d]. Clearly, this leads on to complacency and the privacy risk of becoming accustomed to the *surveillance society*, as discussed below. It is also much easier to accept surveillance when under the influence of someone with whom one has a positive relationship, such as parents recommending their children enable mobile phone location disclosure services [Jiow & Lin 2013].

#### 3.9.1.5 Becoming accustomed to the *surveillance society*

Unfortunately, it seems that it is very easy to forget that one is being observed, as discussed above. This can have two consequences: acting carelessly while being observed and/or accepting the lack of privacy by becoming used to being observed, or even by expecting to be observed. Von Drehle [2013], for example, suggests that Americans have been accustomed to limits on their privacy for many years.

Bentham [1787] conceived of the *Panopticon* as a circular building (for a prison, hospital, workhouse, school or other institution) with an *inspection house* in the middle from which a manager, inspector or custodian could observe the inmates (positioned around the perimeter) without them being aware of when they were being observed or able to communicate with the custodians or other inmates. No true Panopticon prison was ever built, though the Old Provost in Grahamstown used some of the characteristics of a Panopticon. [Foucault 2008] then invoked the concept of the Panopticon<sup>21</sup> as a metaphor for the tendency of modern “disciplinary” societies to observe and attempt to “normalise” their citizens. “*The crowd is abolished. The panopticon induces a sense of permanent visibility that ensures the functioning of power. Bentham decreed that power should be visible yet unverifiable. The prisoner can always see the tower but never knows from where he is being observed*” [SparkNotes Editors nd].

Unsurprisingly, this can lead to very limited, or even curtailed, political and personal freedoms, and even the loss of self-reliance [Carr 2010]. [Dobson & Fisher 2003] took Foucault's metaphor further, identifying three “post-panoptic” models:

1. Bentham's original concept, which they consider to be the one Foucault used;
2. Panopticism II, in the form of the “Big Brother” type of surveillance of Orwell [1949];
3. Panopticism III, technology that tracks humans and their activities, such as cell-phone tracking [Cooper *et al* 2009a, 2010d; Schmitz & Cooper 2011], GNSS receivers, RFID<sup>22</sup> and geo-fences<sup>23</sup>. Crucially, the technology for Panopticism III is relatively

---

<sup>21</sup>Though Brunon-Ernst [2012] suggests that Foucault distorted Bentham's philosophy.

<sup>22</sup>Radio frequency identification, small passive or active transponders.

<sup>23</sup>Virtual or conceptual geographical perimeter or barrier, such as the designated spatio-temporal areas for a parolee with a tracking device.

---

### 3. The context for user-generated content

---

cheap, effective and widely available to anyone, and not just well-resourced national security agencies.

The postal espionage crisis of 1844 in the United Kingdom concerned the opening of Penny Post letters by the Post Office — at the behest of a foreign power, namely the Austrian Empire, wanting to monitor the Italian republican, Giuseppe Mazzini. As the Law Magazine observed in 1845, “the post-office must not only be CHEAP AND RAPID, but SECURE AND INVIOLEABLE” [Vincent 2013]. However, even though the scandal was widely known and caused a ‘paroxysm of national anger’, it did not impact at all on the popularity of the Penny Post, which increased rapidly thereafter [Vincent 2013].

*“Should the global public turn away from the digital media in response to the revelations about the extent of state surveillance, pressure may be generated for reform. Conversely, if the metrics show no decline in use despite all the publicity and debate, conclusions may be drawn which are the reverse of those demanded by liberal protesters. Edward Snowden’s revelations will have demonstrated that in practice, the Web-surfing, texting and emailing public are indifferent to the risks they run to their privacy” [Vincent 2013].*

Similarly, Lanier [2013a] is concerned that 2013 is the year of *digital passivity*, when all the cool gadgets (such as tablets which only run applications approved by some central commercial authority) made us just accept the commercial and government *surveillance economy*. Carr [2010] fears that privacy could become perceived as being an outdated and unimportant concept that inhibits efficient transactions such as socializing or shopping.

#### 3.9.1.6 Mutual surveillance

The psychological and social effects of such prevalent surveillance include self-policing and even being willing to spy on one’s neighbours. These occur in those environments where people are so intimidated by authority and/or so used to surveillance, that they can be forced or encouraged to spy on one another, extending easily, cheaply and significantly the surveillance reach of the authority, be it a government, the military, a corporation or any other type of organisation [Foucault 2008, 1982].

#### 3.9.1.7 Making data already in the public domain more visible

It is a common defence that there is nothing wrong with putting online data already in the public domain, that would otherwise be difficult to access. This would include documents and photographs in archives. However, that does allow data matching, as discussed next. Such online content can be readily accessed by anyone without necessarily revealing their interest in the content, such as using the photography on the likes of *Google Street View* [Google 2016f] to examine a neighbourhood, be it for identifying security weaknesses for targeting burglaries, stalking a resident there, or mere curiosity. Similarly, much data that would otherwise not be available are published, often unwittingly. An example, would be personal details of living people published online in a genealogy.

### 3. The context for user-generated content

---

#### 3.9.1.8 Processing available data

It requires much skill, intelligence and persistence to link together analogue data from diverse sources to find common threads — I was privileged to work with the legendary serial-criminal detective, Piet Byleveld [Schmitz *et al* 2000; Cooper *et al* 2001], who solved many cases by linking together seemingly unrelated clues. Now, massive and persistent digital databases from divergent sources, sophisticated tools (such as for data matching, pattern recognition, behavioural tracking, text analysing, data mining, linkage analysis, statistical analysis, spatial analysis and artificial intelligence), machine translation and fast hardware make combining and analysing data possible, for anyone to identify linkages across data sets that would have been impossible only a few years ago.

Of course, most ‘big data’ analysis is not done to invade privacy, but to allow researchers to examine questions otherwise unexplorable, to understand human, physical and environmental behaviours in different contexts, and (hopefully) create benefits for society in general [Gutmann & Stern 2007]. Unfortunately, it appears that an individual can be identified uniquely with very few data points, and even coarse ones at that, such as with cellular telephone use [de Montjoye *et al* 2013]. It is also possible to identify personal traits from the digital footprints (text, photographs, etc) that people leave on social media and social networking sites, that can also be used for trust and resilience modelling [Zhou *et al* 2013]. Siddle [2014] demonstrates how bike-share data could be used to identify an individual. Hence, Narayanan [2011] states that “there is no such thing as anonymous online tracking”, and Michalevsky *et al* [2015] demonstrates how the mobile telephone’s power consumption alone can be used to determine its location. There are also services available for a fee to track the user of a mobile telephone [Timberg 2014].

“At this time, however, no known technical strategy or combination of technical strategies for managing linked spatial-social data adequately resolves conflicts among the objectives of data linkage, open access, data quality, and confidentiality protection across datasets and data uses” [Gutmann & Stern 2007].

#### 3.9.1.9 Opting in vs opting out

To varying extents in different jurisdictions, one is able to control, to some limited extent, the degree to which one’s personal information is known, retained by others and/or shared. Sharing one’s information (*opting in*) can provide one with access to various services, opportunities or prizes<sup>24</sup>. These might include subscriptions to paid content, exposure of one’s resumé to potential (and hopefully desirable) employers, security services such as vehicle tracking, research collaboration or even friendships or more (such as through dating services). Further, for some *the right of publicity* [Nimmer 1954] is a key part of their profession and their income, through exploitation of their names, photographs, likenesses, recordings and the like — but only if they have consent and are remunerated appropriately. In many jurisdictions, one nominally has the right to *opt out*

---

<sup>24</sup>Which is why there are so many competitions out there, because they are a cheap way to harvest personal data that are up to date.

### 3. The context for user-generated content

---

of divulging one's private information, but even that explicit declaration can get ignored [Harvey 2013b].

Some object to the term *volunteered geographical information* because the information so collected has not necessarily been *volunteered* (ie: the subject opted-in explicitly), but rather has been contributed, collected or harvested irrespective of whether or not the subject opted in, opted out, was not even aware they could be contributing their personal information, or had merely forgotten they were doing so. Harvey [2013b] suggests that we should differentiate explicitly between volunteered (VGI) and contributed (CGI) geographical (or locational) information. Further, he suggests that *truth in labelling* (see Section 6.2) in the metadata that follows pragmatic ethics, would explain the provenance of the information, allowing assessment of its *fitness for use* and determining if the quality of the data has been compromised by lax standards or even malfeasance [Harvey 2013b].

#### 3.9.1.10 Assuming one has nothing to hide

For anyone who lived through Apartheid (or communism, fascism, etc), it should be obvious that everyone probably has something to hide from a repressive government. However, even the courts in a reasonably open and stable democracy such as the United States of America recognise that an innocent person has the right to remain silent [Supreme Court of the United States 2001], and hence keep their matters private. It is not just about keeping the 'facts' about oneself private, but also about the assumptions made about us from the available data [Collins 2014b]. Solove [2007] collected examples of responses one can make to those who justify surveillance, such as: show me the details of your credit-card purchases; show me yours first; it is none of your business; and if you have nothing to hide, then you don't have a life. In any case, the person wanting to protect their privacy does not have to justify their position: the person wanting to invade someone else's privacy needs to justify it first [Solove 2007]. The metadata of one's communications can reveal personality traits, religion, politics, habits, movements, condition, relationship issues, etc [Big Brother Watch nd]. Further, there is the problem of identity theft.

It is appropriate for human beings to have space where they are guaranteed to be free from surveillance or interference by anyone. This applies particularly to establishing and preserving intimate human relationships (eg: dating or dealing with a family crisis), but also for developing intellectual faculties through reading, private conversation or writing privately (eg: in a diary) [Phillipson 2013]. It is very difficult to grow intellectually if one cannot experiment with ideas without fear of surveillance and resulting misinterpretation (or even wear a particular t-shirt [Granick 2013]).

*"Experience should teach us to be most on our guard to protect liberty when the government's purposes are beneficent. Men born to freedom are naturally alert to repel invasion of their liberty by evil-minded rulers. The greatest dangers to liberty lurk in insidious encroachment by men of zeal, well-meaning but without understanding"*  
Brandeis [1928].

Part of the problem is the complexity of legal systems — in the United States, for example, even the Federal Government does not know how many federal crimes there are, which

### 3. The context for user-generated content

---

then also incorporate about 10 000 regulations from various agencies [Marlinspike 2013]. Of course, that also excludes all the offences determined by the legislation of each of the 50 States there. Hence, it is probably easy for innocent people to “commit” serious crimes unwittingly, and for over-zealous law enforcement officials to apply laws more widely than envisaged when they were drafted. A consequence is that the application of the law almost becomes arbitrary. “Sanguinity is misplaced in our current legal regime where so much seemingly innocent behavior arguably fits the definition of one or more crimes” [Granick 2013].

“The skeptics no doubt have noticed that governments are made up of people and that people are prone to misuse information when driven by greed or curiosity or a will to power” [Von Drehle 2013]. Hence, everyone *should* have something to hide.

#### 3.9.2 Censorship

Censorship is not only the suppression of the dissemination or possession of speech, imagery, data or whatever, but also the suppression of the means of access to — or even the desire to try to access — that which is considered offensive, immoral, objectionable, obscene, harmful, embarrassing, threatening, or even (as the opponents of the particular censorship would claim) liberating, by anyone — or even by an algorithm. Censorship is typically done by the powerful (in government, the military, religion, corporations, etc), but can even be done by the weak, in the form of self-censorship or censorship of one’s peers in a marginalised community. Of course, the validity of any rationale for censoring something depends entirely on one’s perspective, and could be quite different for something else.

Censorship is done for many reasons, such as (supposedly) to protect the integrity of the state, in the public interest, to protect the vulnerable (eg: by banning child pornography) and to protect privacy; but also to protect inflated egos (eg: through banning unflattering photographs), to rearrange history (eg: *damnatio memoriae*, ie: trying to erase people from the record; or denying past atrocities), to look after the corrupt and other criminals, to control markets, to safeguard advertising revenues or business, to promote political or religious ideologies, or even just to create confusion [Wikimedia 2016]. For example, President Zuma’s espoused need for good news [Makinana 2013] is also an attempt at censorship, as is his rather hopeful claim that those who insult leaders will be cursed by God [Moloto 2013]. As the Business Day Editor, Zibi [2015], stated, “newspapers cannot provide writers with list of neutral words they may use”. Comedy is susceptible to censorship because if it is not edgy it is boring and irrelevant, but if it is edgy then it is offensive — particularly within the self-righteous ‘Twittersphere’ [Barber 2015].

Censorship can even include censorship of the censorship itself, such as the gag orders obtained by the NSA preventing companies from revealing that their (supposedly private) records of transactions (email, posts, likes, etc) are being searched [Rushe 2013; Hern 2013]. This is now being opposed in court by the likes of Twitter [Bradner 2014]. Censorship is done in a variety of ways, be it through imposing legal penalties, exploiting defamation or slander laws (eg: Mohamed El Naschie’s attack on Nature [Cressey

### 3. The context for user-generated content

2012; Schiermeier 2012)), public burnings of publications, denouncements, peer pressure, blocking Web sites, filtering Internet content (particularly search results, as now required by the European Court of Justice, see Section 3.4.2), super-injunctions in the United Kingdom (used by the famous to try to suppress information about them [Fawkes 2011]), altering maps, doctoring, blurring or blanking out images or text (such as in newspapers), or controlling access to copiers and printers. However, censorship can also be blatant and even implemented as a form of protest by the censored, such as black blocks of ink covering parts or all of newspaper articles.

Censorship can be disguised and rationalized as prudent selection, of course, such as due to the limited budget of a public library, to suppress hate speech, to maintaining literary excellence, to ensure balance and/or to meet the requirements of one's audience [Asheim 1983, 1953]. However, censorship can actually be difficult to implement in a democracy, even one that is not very open. For example, an early version of the Protection of Information Act [South Africa 2010b] allowed for the classification of *categories of information*, which was defined as “*means those groupings, types, classes, file series or integral file blocks of classified information that may be classified, declassified or downgraded together or in bulk*”<sup>25</sup>. This would have allowed information to be declared top secret retrospectively — and the onus would have been on the possessor of such information to know this, understand immediately how it affected their databases, directories, backups, caches<sup>26</sup> and the like, and surrender immediately all such information which they were suddenly not entitled to possess [Cooper 2011a]. Even worse, *information* was defined as “*means any facts, particulars or details of any kind, whether true or false, and contained in any form, whether material or not, including, but not limited to . . . conversations, **opinions**, intellectual knowledge, voice communications and the like not contained in material or physical form or format*” [South Africa 2010b]. Amazingly, the Bill then stated in clauses 4 and 5:

*“1(4) For the purposes of this Act a person is regarded as having knowledge of a fact if —*

- 1. that person has actual knowledge of the fact; or*
- 2. the court is satisfied that —*
  - (a) the person believes that there is a reasonable possibility of the existence of that fact; and*
  - (b) the person has failed to obtain information to confirm the existence of that fact, and ‘knowing’ shall be construed accordingly.*

*1(5) For the purposes of this Act a person ought reasonably to have known or suspected a fact if the conclusions that he or she ought to have reached, are those which would have been reached by a reasonably diligent and vigilant person having both —*

- 1. the general knowledge, skill, training and experience that may reasonably be expected of a person in his or her position; and*
  - 2. the general knowledge, skill, training and experience that he or she in fact has”*
- South Africa [2010b].

<sup>25</sup>The emphases in the quotes here are mine.

<sup>26</sup>To which the possessor could be completely oblivious, such as their cache for a virtual globe that suddenly contained classified data.

### 3. The context for user-generated content

Needless to say, trying to comply with such legislation would have consumed much of any geospatial professional's time! Duncan [2011] described it eloquently as the *Prevention of Scholarship Bill* as it would have outlawed much research and even outlawed identifying empirical gaps in research and examining the research! As discussed above in Section 3.9.1, such legislation would have been implemented arbitrarily because almost anyone would contravene it. This draconian bill was then replaced by the Protection of State Information Bill [South Africa 2013c], and while it got rid of the worst excesses, it still includes clauses 4 and 5, as given above, essentially unchanged. So, given the "general knowledge, skill, training and experience" that I have (admittedly, more than most of the population in all four cases), I "ought reasonably to have ... suspected" that corruption was involved in the arms deal, for example — but how could such a suspicion possibly be removed?

#### 3.9.3 Liability

While geospatial professionals might raise concerns over the quality of VGI and liability for those who use VGI [McDougall 2009], I would suggest that the importance of liability for the quality of geospatial data and services is actually poorly appreciated within the geospatial community in general, with the assumption that all one needs is a disclaimer that the data and services are provided *voetstoets*<sup>27</sup> to protect the supplier from any claim for damages — but this might be inadequate in terms of consumer protection legislation, for example [Rak 2013]. The recent conviction for manslaughter of six scientists and an official in Italy for not predicting the 2009 L'Aquila earthquake [Nature 2012] highlights the risks of not dealing with liability adequately and that the consequences of providing poor data can be serious. Further, while staff of official mapping agencies should generally be aware of such liabilities, this is unlikely to be the case with the public at large: even volunteers contributing data for humanitarian reasons under severe time constraints could still face tort liability [Robson 2012; Coetzee *et al* 2013b].

Uncertainty over liability and privacy can discourage sharing public-sector information (PSI), but having good laws on these can encourage trust and hence sharing [Cooper *et al* 2013]. Further, Foong [2010] suggests that an appropriate Creative Commons licence is sufficient to protect a government releasing PSI to open access, particularly where the limitations of the PSI are made known to the user and "appropriate information management policies and principles are in place to ensure accountability for data quality and accuracy".

Many end users probably cannot differentiate between VGI and official information, unless they are told explicitly, and hence they would use VGI transparently [Cooper *et al* 2011c]. The risks related to using poor quality VGI are essentially the same as those of using poor quality data from an official or commercial supplier [Cooper *et al* 2011a].

Further, even professionals can place undue reliance on digital geospatial data, as happening with the grounding (and subsequent destruction) of the USS Guardian on Tub-

<sup>27</sup>"(Of a sale or purchase) without guarantee or warranty; at the buyer's risk" [Oxford 2016].

### 3. The context for user-generated content

---

bataha Reef<sup>28</sup> in the Philippines on 17 January 2013: the reef had been misplaced by 8 miles on one of the ship's digital nautical charts (DNC), due to inaccurate commercial satellite imagery, but the location was correct on the general DNC<sup>29</sup>. The watch chose to ignore the significance of the flashing light of a lighthouse on the reef and made other mistakes, so the Navy's investigation blamed the crew. Nevertheless, it also made recommendations concerning the DNCs and the user-interface for the navigation system [Haney 2013].

However, there does appear to be increasing demand for conformance testing for, and validation of, geospatial data sets. This would apply especially to dynamic and potentially risky applications such as in-vehicle navigation [Cooper 2013]. Conformance testing and validation assessment are unsurprisingly expensive, because of the expectation of perfection that they can create. Rak [2013] has developed a prototype for validating geospatial data quality and identified four primary risk management techniques specifically for incorporating VGI into official data sets:

1. **Identification of Possible Risks**, such as by forming a risk management team to identify, rank and assess the possible risks, and develop and implement plans to mitigate or even prevent them;
2. **Quality Assurance Procedures**, including comparing to other data sets and getting feedback from users;
3. **Disclaimers in Contract**, developed by legal experts; and
4. Duty to Warn about Quality, both before and after use [Rak 2013].

#### 3.10 The right to exploit content

The right to exploit content is different from the right to control whether or not one's thoughts, sentiments, emotions, likenesses, voice, writings, art works, music or whatever should be made public (the right of an *inviolable personality* [Warren & Brandeis 1890]). These latter rights have been discussed above in Section 3.9.1 on privacy. Different types of ownership of content, or the right to exploit the content, are discussed here. Unfortunately, much content is actually owned by corporations and not the creators of the content and they often have conflicting agendas. This often frustrates the creators — such as the successful South African singer, Toya Delazy, who released her second album online because her record label was not making sufficient stock available in music shops [Channel24 2015].

While a complete work is generally readily identifiable, it can be difficult to distinguish individual units of “intellectual property” (known by some as a *meme*), particularly when they are combined with such units from other creators. Needless to say, this is very common with geospatial data sets, mash-ups and other resources available on the Web. It can

---

<sup>28</sup>Also causing damage to the protected reef, a World Heritage Site.

<sup>29</sup>Where the planned route went right over the reef — and the crew were aware of the discrepancies in the DNCs!

### 3. The context for user-generated content

---

get complex: for example, should a convicted criminal such as the former Panamanian military dictator Manuel Noriega be allowed to get royalties when he is used as a character in a computer game, and should he be able to prevent being portrayed in a negative light [Takahashi 2014]?

#### 3.10.1 Patents

Patent law has been changing recently to increase the powers of patent holders and to extend patent rights into fields such as software [United Nations Economic Commission for Africa 2009]. Unfortunately, it is very expensive to contest patents and it appears that some organisations are using them as a form of a trade barrier. Stallman [2008a] suggests that “programmers are well aware that many of the software patents cover laughably obvious ideas”, and proceeds to dissect a software patent he considers to be trivial and for which he considers there was prior art. Dubious patents from 2015 include one for changing the quantity of goods being ordered, a firewall that cannot be configured, connecting anything to the Internet and using electronics to control a sex toy [Downes 2015]! Even for the holder of the patent, their rights can be fragile and limited, with issues such as territoriality, non-examination in jurisdictions such as South Africa, the scope of protection, enforceability, and the fees for prosecution and renewal [Biagio 2013].

Stallman [2008b] also considers it inappropriate to lump together all of the legal mechanisms for patents, copyright, trade marks, trade secrets and the like under the label *intellectual property*, as he considers it to be a “distorting and confusing term that did not arise by accident”, because it tries to make them analogous to property rights for physical objects. There are significant differences between patents, copyright, trademarks, trade dress, designs, geographical indications (eg: appellations of origin), new botanical varieties, trade secrets, and the protection of performances, which are then blurred by giving them all a collective label such as “IPR”. This conflates “rival” goods, which cannot be shared without cost, with “non-rival” intellectual products that can be shared simultaneously by everyone [Stallman 2008b]. Further, there is the false presumption that IPRs always promote innovation [Mallaby 2012]. The patent system has been a pillar of the American capitalist vigour, but it has become decadent with the plethora of suits over alleged theft of intellectual property — “American law is patent nonsense” [Mallaby 2012]. The result is patent trolls and the like.

The United States Patent and Trademark Office (USPTO) has a backlog of about one million patents and their patent examiners have only about 20 hours to examine each patent application and to determine whether or not it deserves a patent [Allen *et al* 2008, 2009]. With the New York Law School, various information technology companies and foundations, the USPTO set up a pilot project, *Peer-to-Patent*, which ran from 15 June 2007 to 15 June 2009 and which opened the patent examination process to public participation to help the USPTO find relevant information for assessing the claims [Peer to Patent 2016].

The original purpose of awarding patents was valid, but over the centuries the environment has changed and given the extent to which patents are abused now, the patent system needs radical overhaul. Elon Musk has now made all the patents of his compa-

### 3. The context for user-generated content

---

nies such as Tesla, open to stimulate innovation: “When I started out with my first company, Zip2, I thought patents were a good thing and worked hard to obtain them. And maybe they were good long ago, but too often these days they serve merely to stifle progress, entrench the positions of giant corporations and enrich those in the legal profession, rather than the actual inventors. After Zip2, when I realized that receiving a patent really just meant that you bought a lottery ticket to a lawsuit, I avoided them whenever possible” [Musk 2014].

#### 3.10.2 Copyright

Copyright is not just for giving the creator of an original work (however that might be defined) the exclusive right to exploit the work (particularly, produce and sell copies), but also for ensuring that the creator is appropriately credited and that the owner of the copyright can also decide who can adapt, perform, broadcast or otherwise benefit from the work, and how. The ownership of a copyright may be sold<sup>30</sup>, a copyright generally expires after a fixed term and in most jurisdictions (those adhering to the Berne Convention [WIPO 1979]), the copyright is awarded automatically without the need for formal registration. There are also variations over what may be copyrighted, such as the expression of an idea but not the idea itself, the aesthetic features but not the utilitarian functions of a useful object, facts, and fair use [Wikimedia 2016]. As with patents, some claim that copyright criminalises legitimate use in enforcing the interests of large corporations.

When the copyright expires, the work enters the *public domain*, whereupon it may be used or exploited freely by anyone. This is not the same as being *publicly available* (such as over the Internet), though some people do confuse the two. There are also alternatives to standard copyright for facilitating the legal sharing of works, the best known of which is *Creative Commons* (CC), with six general types of CC licences that anyone can use for free, that specify which rights have been waived by the creator. Essentially, the six CC types are standard contracts [Creative Commons 2016]. For software, Creative Commons recommends licences specifically for software, such as the GNU General Public License (GPL) [Free Software Foundation 2016].

Copyright has been a long-running issue with geospatial data, particularly concerning the maps produced by national mapping agencies that were then digitized by other organisations (eg: companies, universities or even other government departments) and distributed in whole or piecemeal, as raster scans or vectorized data, and possibly combined with other data sets. The quality of the digitizing could vary from superb to abysmal, but the result would invariably be described as being data from the relevant NMA, much to their chagrin. This would contravene the creator’s rights to preserve the artistic and scientific intent of their works (the maps) as well as cast aspersions over the quality of the maps, never mind deprive them of possible income.

The problem has not gone away with the distribution of digital data sets by NMAs, nor with legislation making public-sector data available without charge, such as South Africa’s Promotion of Access to Information Act [South Africa 2000] and Spatial Data Infrastructure Act [South Africa 2003]. Hopefully though, the greater awareness and

---

<sup>30</sup>There are many sad cases of this having been exploited by the ruthless.

### 3. The context for user-generated content

---

proper use of metadata will make the provenance of geospatial data more obvious to the end users.

#### 3.10.3 Open access

There are different perspectives on what it means to be “open”, and different degrees of openness. The Open Definition [2016] provides a summary definition of *openness* as being: “A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or sharealike”. Their fuller definition also deals with issues such as using open (non-proprietary) formats and not discriminating against, people, groups of fields of endeavour in granting access.

There are authors, such as Boyle [2007], who have realised that encouraging free (non-commercial) copying of their works can create a bigger market for their books (creating awareness rather than obscurity) and related services, such as consulting and speaker fees. Even as famous an author as Paulo Coelho supports the file-sharing site The Pirate Bay, because he feels that it increases sales of his books [Wikimedia 2016].

More importantly, many governments understand the importance of making public-sector information (PSI) readily available (such as the European Union’s Directive on the reuse of PSI [European Parliament 2003]). This Directive aims to reduce duplication within government, to improve service delivery and to stimulate economic activity through exploiting such PSI, be it to improve the products and services of a company, or to embed the PSI in a mash-up or geospatial data set, or to analyse the PSI to produce new insights. However, the message does not necessarily sink in across government. As Gray [2009a] highlighted, South Africa’s Intellectual Property Rights from Publicly Financed Research and Development Act [South Africa 2008a] contradicts South Africa’s support of the World Health Organization’s resolution of 22 May 2009, for collaboration on world public health through “open-source development, open access to research publications and data, voluntary provision of access to drug leads, open licencing, and voluntary patent pools” [Gray 2009a] (which aligns with the Promotion of Access to Information Act [South Africa 2000]).

Unfortunately, it would appear that the IPR Act was motivated and written by bean counters with no experience of actually creating intellectual property of any value. For example, the definition of *intellectual property* in the IPR Act “*excludes* copyrighted works such as a thesis, dissertation, article, handbook or any other publication which, in the ordinary course of business, is associated with conventional academic work” [South Africa 2008a]. Further, the Act requires disclosure by researchers of all *possible* intellectual property within 90 days.

There are many opposed to open access to data. The Sunlight Foundation has collected over 50 excuses (predictable, political, personal, practical, etc) for denying open access to data, such as staff training, privacy, security, difficulty, costs, liability, data quality, the data are lost, lack of interest, or even that it is unnecessary because the data are already available [McCann 2013]. Needless to say, the Sunlight Foundation has also been compiling rebuttals of these excuses [McCann & Green 2013]. Of course, there are data that

---

### 3. The context for user-generated content

---

should not be released, “but this needs to be a confident determination, not the default option for all data” [McCann 2013]. van Genderen [2013] conducted a comparative survey of legislation dealing with access to PSI in India, Indonesia, The Netherlands and the United Kingdom, concluding that there are striking differences between developed and developing countries in their implementations of freedom of information legislation. She found that while the two developing countries had good FOI legislation, they still have some way to go to implement the legislation successfully and to inform the public of their rights and they are constrained by a secrecy culture, mismanagement, corruption and illiteracy.

Slee [2013] cautions that one needs resources to be able to exploit the availability of open PSI and other resources, with “the emergence of a winner-take-all economy in which small organizations and small businesses are severely handicapped against those with capital behind them”. Hence, there needs to be some form of protection for civic-scale organisations against those with the scale to deliver efficiency, but not participation [Slee 2013]. Lanier [2013b] suggests that one should be paid for all that one contributes (including one’s likeness on CCTV cameras in public) — to monetise data and information. This would make government spying more efficient by reducing the benefit of wasting money on data of innocent people. More importantly, he suggests that it would boost the global economy, because “in a world of free information, the economy will start to shrink as automation rises radically. This is because in an ultra-automated economy, there won’t be much to trade other than information” [Lanier 2013b].

#### 3.11 Curation

*“The massive increase in digital information in the last decade has created new requirements arising from a deficit in the institutional and technological structures and the human capital necessary to utilize and sustain the abundance of new digital information. ... Digital curation differs from traditional curation of physical objects and collections because of the dynamic nature of digital information, its dependence on hardware and software for processing and analysis, its fragility, and many other characteristics. ... Furthermore, because digital information is fragile, corruptible, easily altered, and subject to accidental and intentional deletion, maintaining the integrity of information is a critical aspect of digital curation. Digital curation can enhance the integrity of digital information and increase its trustworthiness through security and restricted access to curation systems, replication, documentation of any transformations of the information, and auditable process and procedures”* [National Research Council 2015b].

This study by the National Research Council in the USA found that digital curators fall into different occupational categories, and there is a poor match between the knowledge and skills needed and the existing jobs [National Research Council 2015b]. Vines *et al* [2013] have discovered that the availability of research data to support the findings in articles in the literature declines rapidly with the age of the article, at a rate of about 17% per annum, with the key problems being invalid email addresses and obsolete storage

### 3. The context for user-generated content

---

technologies. Uhlig [2010] identifies three key issues concerning for the preservation of scientific data:

- The value of data increases with their use;
- Public and publicly funded information want to be free; and
- Digital resources will not survive or remain accessible accidentally [Uhlig 2010].

Projects such as the *Data-At-Risk Initiative (DARI)* aim at understanding the problems and taking action because “valuable and unique scientific data are increasingly at risk of being lost forever due to deterioration, format obsolescence, and insufficient metadata for discovery and retrieval [Murillo *et al* 2012]. Currently under development within ISO/TC 211 is ISO 19165, *Geographic information – Preservation of digital data and metadata*, with the scope to set “the rules for the long-term preservation of digital geospatial data. These data include metadata and other ancillary data that are necessary to fully understand and rebuild the archived digital environment”. ISO 19165 draws on ISO 14721:2012, *Space data and information transfer systems — Open archival information system (OAIS) — Reference model* and related standards.

It should be borne in mind that these archives need to be preserved and made accessible not only for the use of the material in contemporary research, but also for historical research in the future. Unfortunately, digital archives are also far less robust than paper-based archives, so digitizing them is a decidedly complex issue [Cooper 2009b]. On the other hand, even potentially valuable book collections are being destroyed in massive amounts by librarians in developed countries, such as Australia [Waterford 2013].

Without archiving in repositories that are readily available, research and scholarship can easily become invisible: “as supply tends to create demand, so too, absence of supply tends to cause absence of demand” [DG Kourie 2014, *pers comm*]. This can be a particular problem in developing countries with the lack of resources, such as in Africa. Hence, “we are probably the only part of the world about which it is still legitimate to publish without reference to local scholarship” [Mkandawire 1997], which is driven by “the failure to establish intellectual bridges ... and the invisibility of African scholarship”. More than a decade later, the problem has not changed. “You only have to look at the bibliography section of any article written by a western Africanist on Africa and you hardly see any reference made to African sources. No reference to any published works by African scholars and published inside the continent — as if there are no scholars in the entire continent” [Cheru 2012]. This problem is not limited to the social sciences, of course.

Much of the information archives in Africa are still paper-based (including maps), which limits their accessibility but does not mean that they are not very useful. Mkandawire [1997] also suggested that many African universities had empty libraries, which makes it even more difficult to access the literature. This has been borne out by a research project we have with the University of Namibia<sup>31</sup>, where our collaborator has online access to only the most recent couple of years of issues of relevant journals. Sturges & Neill [2004] describe this as the *quiet struggle*, the struggle for information and knowledge by the

---

<sup>31</sup>Modelling a National Health Spatial Data Infrastructure for Namibia, funded by the South African — Namibia Research Partnership Programme Bilateral Agreement.

### 3. The context for user-generated content

likes of “writers, journalists, publishers, educationalists, broadcasters, media workers, computer and telecommunication specialists and, not to be forgotten, its information and library professionals” Sturges & Neill [2004].

As I have stated before [Cooper 2009b], key resources that should be digitized and put online are listed below.

- Issues of **African academic journals** and books that are not yet available digitally. Because they are not digital and not online, they don’t get picked up by search engines such as Google Scholar<sup>32</sup> [Google 2016e], hence reducing the likelihood of African research being cited and used by other researchers, and hence diminishing the value of research done in Africa.
- **Proceedings of African conferences.** The situation of proceedings is even worse than it is for journals, as many proceedings in the past will have been published without International Standard Book Numbers (ISBNs) and hence will not have been lodged in the relevant legal deposit libraries. They might also only exist in private collections.
- **Project reports and data sets** that are unlikely to form part of national archives. This is particularly important for longitudinal studies (many of which are now citizen-science projects, such as the Second South African Bird Atlas Project) and for other data sets that can be used for any long-term studies, such as climate change. African user-generated content submitted to global projects, such as Wikipedia, OpenStreetMap or Tracks4Africa, will be curated in common with those data holdings for other parts of the world.
- Collections of photographs, films, videos and audio recordings.
- Photographs and documentation of **tangible objects** of scientific or cultural value worth preservation, such as historic scientific equipment (including computers), specimens, cultural artefacts and original manuscripts.

#### 3.12 The digital divide

The term *digital divide* is used to highlight significant differences in access to computing and telecommunications within and between communities, countries and regions, and even between ideologies or political systems: the power asymmetries within cyberspace. It is not only about technology, but also about the ability to harness the technology, whether for legitimate or illegitimate reasons [Burrows *et al* 2012]. Guillén & Suárez [2005] argue that democracies promote faster growth of Internet use because it is a threat to authoritarian or totalitarian governments, hence increasing the digital divide between them.

It is trite to consider the digital divide to be one-dimensional (as is sometimes the case), because there are many different underlying aspects: costs, bandwidth, network relia-

<sup>32</sup>Which also do not access the deep Web.

### 3. *The context for user-generated content*

---

bility, electricity supply reliability, censorship, peer pressure, and cultural and linguistic barriers or enablers. For example, fixed-line telephony in Africa has always been poor — but this meant that the roll-out of wireless telephony would not compromise significant investments in fixed-line telephony infrastructure, and to some extent would allow African societies to “leap frog” ahead of some more developed countries that were constrained by their investments in fixed-line infrastructure. What has facilitated the deployment of the base stations for cellular telephony, is that each base station can be a self-contained island with its own electricity generating capacity and security, making it completely independent on the often unreliable electricity supply infrastructure in many African countries. By early 2013, there were 6.8 billion mobile device subscriptions for a global population of 7.1 billion [ITU 2013*b*]. Yet, as recently as 2008, the novelist and Nobel Laureate, JMG le Clézio, said in his Nobel Lecture: “to provide nearly everyone on the planet with a liquid crystal display is utopian” [Le Clézio 2008]!

In many parts of the world, assuming first that a potential user actually has access to electricity and a computer, fixed-line Internet access is very slow, very expensive and unreliable, with fixed-broadband penetration of only 6% in the developing world. Less than 10% of those fixed-broadband subscriptions in Africa offer speeds of 2 Mbit/s or more [ITU 2013*b*]. Because of the poor service offered for land lines in South Africa, I and many other users have wireless access at home as our primary means of access, which in 2010 typically yielded less than 5 Mbit/s [MyBroadband 2010]. Unsurprisingly then, South Africa’s fixed-telephone subscriptions per 100 inhabitants are declining (8.2 in 2011 to 7.9 in 2012) while mobile-cellular subscriptions already exceed the total population and are still increasing, from 126.8 per 100 inhabitants in 2011 to 134.8 in 2012 [ITU 2013*a*]. The South African government only expects to have universally accessible broadband across the country by 2030 [Dawood 2014*a*].

On average, mobile broadband is extremely expensive in Africa, at 38.8% of gross national income (GNI) per capita for a 1 GB plan, which is more than six times the cost in any other region, and more than 32 times the cost in Europe! Nevertheless, mobile broadband is still significantly cheaper than fixed broadband in Africa, at less than three-quarters the cost of fixed broadband, on average. Mobile penetration in Africa grew rapidly from 2% in 2010 to 11% in 2013 [ITU 2013*b*], and 73% in 2015 [World Bank Group 2016]. Now, more households in developing countries have mobile telephones than have electricity or improved sanitation [World Bank Group 2016]. In 2014, the Chair of the Federal Communications Commission (FCC) in the United States of America, pointed out that while 80% of American households had broadband access of 25 Mbps or better and half had access of 100 Mbps, as the bandwidth increases in the USA, the competitive choices decrease. He also considered it unacceptable that 40% of American households did not have access to 100 Mbps broadband [Wheeler 2014].

This growth of mobile subscriptions in developing countries provides opportunities for all kinds of online information and applications, including geospatial data and services, such as geovirtual environments (virtual globes and the like, see Section 2.10). Access to mobile networks is now available to 90% of the world’s population and even to 80% of the population living in rural areas. There is also a rapid move from 2G to 3G platforms, in both developed and developing countries. In 2010, 143 countries were offer-

### 3. The context for user-generated content

ing 3G services commercially (including most of sub-Saharan Africa), compared to 95 in 2007. Many (but not all) 3G systems provide mobile broadband access, but there is sometimes a premium that one pays for 3G access, which then renders it more expensive than fixed broadband [ITU 2010; Cooper *et al* 2011b]. Data-enabled smartphones are not yet common in Africa, which is why Wikipedia, for example, is testing an “article via SMS” service in Kenya [Lee 2013b]. However, the growth of mobile telephony can be constrained by the lack of know-how and money for making available more spectrum, such as through auctions [Reuteurs 2012].

Nominally, the *digital divide* is caused by the lack of access to capital, the legacy of colonialism and factors such as limited personal wealth, education and skills — though initiatives such as the CSIR’s *Digital Doorway* [Gush *et al* 2004] and its predecessor in India, the *Hole in the Wall* experiment [Mitra & Rana 2001], have shown that skills and education are actually **not** barriers to using information and computing technologies. The diffusion of innovations, such as the Internet, World Wide Web and associated services, are determined by the economic, political and sociological context, as well as by the merits of the technology — but having an enabling infrastructure that is affordable is critical [Guillén & Suárez 2005]. In some cases, the digital divide is really due to the governments of the less developed countries, mainly for *pseudo-security* and because of an *enforced telecommunications monopoly*. These are discussed below, but see also the discussion above on attempting to control the Internet, in Section 3.8.

#### 3.12.1 Enforced telecommunications monopoly

Across Europe, the presence of monopolistic markets has been a major inhibitor for providing network access. The problem has never been a lack of technical knowledge, but restrictions imposed artificially by the market. The average price for bandwidth in Europe in 2002 was €5 000 per Mb/s per year, but in European countries without monopolies it was as low as only €36 per Mb/s per year — and the divide was widening [Stöver 2002]. Hence, there was an emerging digital divide being created in Europe, between the countries effectively with open networks, and those with restrictions. Bandwidth costs had been cut by a factor of over 6000 over five years then, but there was also a very large divergence between the cheapest and average costs of bandwidth [Davies 2002]. Not only does competition reduce the costs of telecommunications, but it also improves service differentiation, choice and quality — and increases the marketing of the services [Guillén & Suárez 2005].

#### 3.12.2 Pseudo-security

Some governments use archaic laws about national security to deny their citizens access to data, especially geographical data such as large scale maps or even the use of GNSS receivers (as was the case in Egypt until April 2009 [Privat 2009]). The irony is that their “enemies” probably have much better data readily available (higher resolution, more current, digital as opposed to analogue, with decent metadata, etc), and make such data

### 3. The context for user-generated content

---

available to aid agencies, non-governmental organisations (NGOs), private companies and the like.

This is the excuse, of course. The real motivation for denying access to such data by their citizens, is that it will expose the inadequacies of the governments, revealing their inability to deliver services — and especially, their corrupt practices, with an example being attempts to control all information about the results of the recent elections in Zimbabwe [Roper 2013]. Democratisation of data (see Sections 4.3.4.1 and 4.5.2.1) is a crucial component in enabling the citizens of a country to rescue their country from the politicians; it ensures that everyone works from a common basis and understanding of the realities of the country and region, reducing unnecessary conflict over understanding what the problems are that need to be addressed. “Democratic political regimes enable a faster growth of the Internet than authoritarian or totalitarian regimes, controlling for economic development and income”, because media enabling decentralized or public mass communication (as the Internet does) “undermine the effectiveness of authoritarian or totalitarian rule by allowing citizens to secure their own information ... and to communicate with one another and potentially mobilize politically” [Guillén & Suárez 2005]. Faris & Villeneuve [2008] identified four reasons for filtering the Internet (that is, enhancing the digital divide by suppressing content and services): politics and power, social norms and morals, security concerns, and Internet tools (including VOIP as a commercial threat).

#### 3.12.3 The size of the digital divide

The size of the digital divide is illustrated by the Digital Access Index for 2006 from the International Telecommunications Union (ITU), which showed that although there were 57.3 Internet users per 100 inhabitants in Sweden, 57 in the United States, and 34.7 in Italy, there were just 0.5 in Mali and 0.2 in Niger. The Internet tariff for the same type of connection then was 1.1% of the Gross National Income in Sweden and the United States and 1% in Italy, whereas it was 289% in Mali and 683% in Niger [Zennaro *et al* 2006]. With growing bandwidth requirements (due to increasing numbers of users, increasing use of bandwidth-hungry applications, and expectations of general broadband availability by application developers), many Internet applications became increasingly impossible to use in many universities and organisations across Africa [Zennaro *et al* 2006].

While Internet access is improving in Africa, Africa is not necessarily catching up to the rest of the world — many areas could still be falling further behind as Internet access improves even more quickly in developed countries. The recent installation of several undersea cables (eg: EASSy and SEACOM) should improve the situation, though the South African government tried to block them landing in South Africa because of ownership requirements and other reasons [Vecchiatto 2007; Malakata 2006]. Broadband access in South Africa more than doubled between 2010 (3.6 million subscriptions) and 2012 (8.2 million) [WorldWideWorx 2012].

#### 3.12.4 Research issues

Drawing on Cooper *et al* [2011b], possible research issues concerning the digital divide and VGI, geovirtual environments and SDIs include:

1. **Information poverty and the digital divide.**

Research how SDIs, geovirtual environments and other repositories of geospatial data (whether accessed through mobile devices computers) can address *information poverty* and the digital divide. “In many parts of the developing world, poverty is exacerbated by information poverty. In poor or deprived communities access to information is limited or non-existent” [Pandor 2010].

2. **Virtual globes.**

Do virtual globes and other repositories of VGI entrench the digital divide, because the better resourced are able to provide more data about their home turf? Does the bandwidth available in better resourced areas encourage the people there to contribute more VGI? for example, there are more articles in Wikipedia about fictional places such as Middle Earth and Discworld, than there are about many real countries [Graham 2009]. As Le Clézio [2008] cautions, “are we not, therefore, in the process of creating a new elite, of drawing a new line to divide the world between those who have access to communication and knowledge, and those who are left out?”

3. **Mobile devices.**

Develop novel and cheap ways of representing SDIs, geovirtual environments and other repositories of geospatial data on mobile devices with small information displays and with limited bandwidth.

4. **Spatial context.**

Research how delivering geospatial data to mobile devices can help users (both literate and illiterate) understand their spatial surroundings and how their actions can affect others — the world is bigger than one little village or suburb. An example could be showing how a river still has to support communities downstream, whether one is in an urban or a rural environment.

5. **Bandwidth.**

Does too much bandwidth actually result in lower-quality VGI, effectively providing quantity rather than quality? In other words, if it is expensive for someone to contribute VGI, do they pay extra care to the quality of their VGI?

6. **Display priority.**

How should a virtual globe decide how to prioritise the data that can be displayed, not just to enable rapid zooming in and out of the background imagery, but also to reduce visual clutter without hiding critical information? Determining how to prioritize what is shown on top of other data (and hence obscuring them) is not a neutral process, because of the political and other implications [Cooper *et al* 2011b].

7. **Other issues.**

In the context of the digital divide and VGI, geovirtual environments and SDIs,

### 3. The context for user-generated content

---

Research issues such as the ownership of the data, quality assurance (particularly of the VGI), anonymous contributions, the political and other agendas embedded in the data, and facilitating or denying access to the data.

## 3.13 Standards

### 3.13.1 Overview

A standard is a “document, established by consensus and approved by a recognised body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context” [ISO/IEC 2004]. ISO’s Concept Database contains all the terms defined in ISO’s standards and actually has 59 entry for ‘standard’, those these either use this definition above or are specifically for their domain.

Unfortunately, there is a perception that standards merely encode existing technology and hence that there is no research or innovation in standards development. As a result, academics at South African universities (as is the case in many other countries) do not get credit for their involvement in writing standards, whereas they do for publishing in the peer-reviewed literature: yet a standard undergoes a much more rigorous review by many more experts than is the case with a journal article. There is actually a synergistic relationship between research, innovation and standardization [Coetzee & Cooper 2012].

As discussed above in Section 2.2.3.5, the primary source for standards for geospatial data is ISO/TC 211, *Geographic information/Geomatics*, for which the local mirror committee at the South African Bureau of Standards is SABS/TC 211. ISO/TC 211 has 38 participating (P) members and 29 observing (O) or corresponding members. From Africa, Botswana and South Africa are P-members, Kenya, Mauritius, Morocco and Tanzania are O-members and Swaziland is a corresponding member. ISO/TC 211 also has liaisons with 32 organisations, including key standards generating bodies such as the Open Geospatial Consortium, Inc (OGC), the Defence Geospatial Information Working Group (DGIWG), the International Hydrographic Organization (IHO), the Scientific Committee on Antarctic Research (SCAR) and the Universal Postal Union (UPU). On behalf of ISO/TC 211, the OGC and the IHO, [Bessero *et al* 2013] compiled a report for the United Nations initiative on Global Geospatial Information Management (UN-GGIM), providing recommendations on setting standards in the global geospatial community to support the aims and objectives of UN-GGIM, such as developing, aggregating and arranging global geospatial information and promoting access, reuse and its application for addressing key global challenges.

Standards developed by SABS/TC 211 include **SANS 1878-1:2005**, *South African spatial metadata standard, Part 1 — Core metadata profile* [SANS 1878 2005], which is the South African profile (sub set) of ISO 19115:2003, *Geographic information — Metadata*; **SANS 1876**, *Feature instance identification standard* [SANS 1876 2013] (which is still a draft); **SANS 1880:2014**, *South African geospatial data dictionary (SAGDaD) and its application* (for which I

---

### 3. The context for user-generated content

---

was the Project Leader); and **SANS 1883-1:2009**, *Geographic information — Addresses Part 1: Data format of addresses*.

Unsurprisingly, there are those who are sceptical about standards. Swartz [2013] suggests that standards should be written after one has got something to work, not before, citing the example of JSON (which he considered a sensible format) against XML (which he considered to be a scourge on the planet). Developing data models for key data sets could help develop a culture of standards and best practice in an organisation [Hughes 2005]. Standards underpin quality (see Chapter 6), metadata (see Chapter 5) and taxonomies (see Section 2.4), and come out of experience with them.

#### 3.13.2 Standards for volunteered geographical information

In general, contributions to a repository of VGI need to conform to various standards specified for the repository, particularly concerning the structure or syntax of the data. These are often proprietary standards (eg: OSM XML for OpenStreetMap [2016]), though they might incorporate aspects of widely used standards, such as those from ISO/TC 211 [ISO/TC 211 2016], or the Open Geospatial Consortium (OGC) [OGC 2016]. Further, there are often standards for the content of the data, such as for citizen science projects that specify a protocol for recording the data in the field (eg: Animal Demography Unit [2016b]). Unfortunately, it is difficult to assess adherence to such a protocol, particularly as some citizen scientists might over estimate their adherence to the protocol<sup>33</sup>. My experience with SABAP2 is discussed in Section 6.8.1.

A further complication is that in general, users are not involved in the development of standards, such as for assessing quality or documenting metadata. The result is that even if they are aware of the relevant standards, they do not necessarily “buy in” to the standards nor understand their context or utility. Additionally, in our experience, even GISc professionals can struggle to read a standard without some training, because of the formal requirements for a standard and the necessarily repetitive structure of the text — a standard is not a novel [Cooper *et al* 2011a]!

## 3.14 Summary and looking ahead

This chapter has provided details of the context that made the proliferation of user-generated content and volunteered geographical information possible, and the impact thereof: inter-networking, services, the Semantic Web, social mapping, (impossibility of) controlling the Internet, open archives and access, privacy, censorship, liability, patents, copyright, curation, the digital divide and standards. All of these impact on VGI and SDIs. While this chapter provides the setting for subsequent chapters, it also makes important contributions as part of my research and this thesis.

---

<sup>33</sup>For example, I recall from my days as a cricket umpire, how few cricketers knew the Laws of Cricket [Marylebone Cricket Club 2010], even those who were experienced first-class players.

### *3. The context for user-generated content*

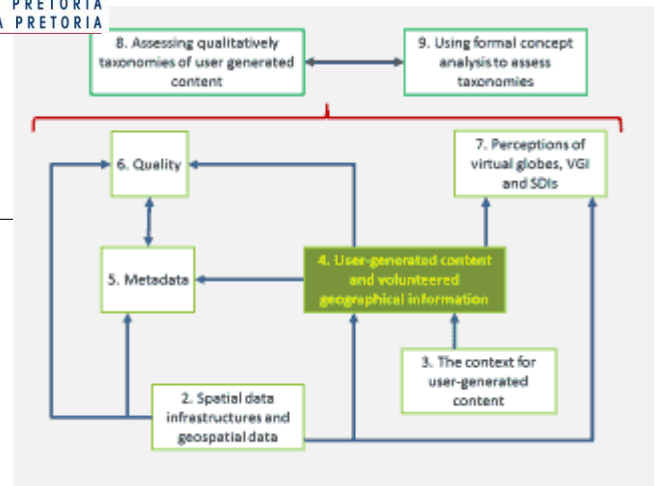
---

Chapter 4 builds on this chapter and Chapter 2, and will now provide details of user-generated content, citizen science, volunteered geographical information, crowd sourcing, neogeography, the validity of using user-generated content in scholarly research, the quality of the traditional scholarly media, and citing user-generated content, data and data repositories.

\*\*\*\*

---

### *3. The context for user-generated content*



## Chapter 4

# User-generated content and volunteered geographical information

### 4.1 Overview of the chapter

Chapter 2 discussed spatial data infrastructures (SDIs) and geospatial data and Chapter 3 provided the context for user-generated content in general. This chapter draws on them to present a detailed discussion of user-generated content, citizen science, volunteered geographical information, crowd sourcing and neogeography, which are often confused with one another: see Section 4.2 and Table 4.1 for examples of differences. Further, this chapter discusses the validity of using UGC in scholarly research, the quality of the traditional scholarly media, how traditional scholarly media matches official producers of geospatial data, and citing UGC, data and data repositories. Specifically, this chapter covers the following.

- Section 4.2 introduces user-generated content, crowd-sourcing, citizen science and neogeography, and explains why *they are not the same*.
- Section 4.3 discusses *user-generated content*, including definitions, the nature of historical and modern UGC, and aspects of UGC such as data democratization, freedom of information, control of data dissemination, invasion of privacy, abuse, involuntary UGC, the digital divide, anonymous contributions, the loss of context, broken links, quality, proof by repeated assertion and bias.

#### 4. User-generated content and volunteered geographical information

---

- Section 4.4 discusses *citizen science*, covering the nature of the interactions between science and ordinary citizens: science communication, educating the public-at-large, using sound science and contributions by the public-at-large to science, known as citizen science. This section discusses and extends the typology of citizen science of Wiggins & Crowston [2011], which is used to assess VGI repositories in Sections 8.5 and 9.6.6. Finally, this section provides an overview of potential problems with citizen science.
- Section 4.5 discusses *volunteered geographical information (VGI)*, covering the various interpretations of what VGI actually is and aspects of VGI: these are obviously similar to the aspects of UGC in general discussed in Section 4.3, with the focus here on the geospatial context, but this section also discusses contributing to and competing with an SDI, technologies, motivations of contributors (discussed further in Chapters 8 and 9), types of VGI (also discussed further in Chapters 8 and 9), mis-registration of VGI and metadata.
- Section 4.6 discusses *crowd sourcing*, addressing the nature of crowd sourcing (which seems to be misunderstood by some) and types of crowd sourcing, such as targeted crowd sourcing, competitions, surveys, crowd funding, open review, open design, the (alleged) wisdom of the crowd or collective intelligence, micro volunteering and open source software. It also assesses the crowd-sourcing taxonomy of Saxton *et al* [2013].
- Section 4.7 discusses *neogeography*, a term with various interpretations, though it should just be considered broadly to cover new ideas, theories, sub-fields and applications in geography (which is what the word really means). This section covers the historical and contemporary perspectives of neogeography and raises a concern over being distracted by the anti-intellectualism aspects of some post-modernists and the like.
- Section 4.8 discusses the *validity of user-generated content in scholarly research*, covering the quality of UGC, blogs and the quality of the traditional scholarly media, including threats to science, the publish or perish dilemma, traditional scholarly journals, sloppy editing or refereeing, quality of citations, cliques, gurus, obsequiousness, predatory open-access publishers, unethical authorship practices, fake science, fake journals, unethical practices by publishers, plagiarism and the decline effect. This section also compares traditional scholarly media with official producers of geospatial data (as UGC is to traditional journals, so is VGI to SDIs) and discusses what might lie beyond traditional scientific media.
- Section 4.9 discusses *citing* UGC, data and repositories.

The major original contribution that I have made that is presented in this chapter is:

- To information science, clarifying the differences between user-generated content, volunteered geographical information, citizen science, crowd sourcing and neogeography, because it appears that they are often confused with one another. See Sections 4.2 to 4.7, inclusive.

Further key contributions that I have made that are presented in this chapter are:

#### 4. *User-generated content and volunteered geographical information*

---

- Identifying the nature and various aspects of user-generated content, see Section 4.3.
- Expanding on the typology of citizen science of Wiggins & Crowston [2011], see Section 4.4.2.
- Identifying various aspects of volunteered geographical information, see Section 4.5.2.
- Identifying types of crowd sourcing, going beyond the commercially-oriented taxonomy of Saxton *et al* [2013], see Section 4.6.3.
- Identifying the types of blogs, and how these types correlate to volunteered geographical information, see Section 4.8.2.
- Identifying problems with assuming that traditional scientific media are inherently of a high quality, see Section 4.8.3.

This chapter also answers questions such as:

- What is UGC, VGI, citizen science, crowd sourcing and neogeography?
- Why are citizen science, crowd sourcing and neogeography not restricted to user-generated content?
- Why do professionals also have a role in neogeography?
- Can user-generated content be used legitimately for scholarly research?
- What lies beyond traditional scientific media?

Sánchez-Vaquerizo *et al* [2015] state provocatively:

*“The current large availability of data recently published by public agencies is absolutely wasted. Open big data remain useless if they are not visualized and interpreted. Public institutions are just worried about getting involved in the current public demand of transparency and accountability oriented to good governance, but only in a superficial way: keeping up appearances. Facing their lack of interest for truly broadcasting of those data, collaborative working and mapping collective online tools are powerful resources for citizens for making the most of those supposedly uninterested published data”* [Sánchez-Vaquerizo *et al* 2015].

## 4.2 User-generated content, crowd-sourcing, citizen science and neogeography are not the same!

The concepts of volunteered geographical information (or user-generated content), crowd-sourcing, citizen science and neogeography are sometimes confused with one another. While they can overlap, each has its unique characteristics. The following are dictionary definitions of the terms, with *new media* added for context.

- **New media:** “content available on-demand through the Internet, accessible on any digital device, usually containing interactive user feedback and creative participation” [Wikimedia 2016].

#### 4. User-generated content and volunteered geographical information

- **User-generated content:** “denoting or relating to material on a website that is voluntarily contributed by members of the public who use the site” [Oxford 2016]. This definition is rather limited when compared to that of Wunsch-Vincent & Vickery [2007], see Section 4.3.1.
- **Crowdsourcing:** “obtain (information or input into a particular task or project) by enlisting the services of a large number of people, either paid or unpaid, typically via the Internet” [Oxford 2016].
- **Citizen science:** “the collection and analysis of data relating to the natural world by members of the general public, typically as part of a collaborative project with professional scientists” [Oxford 2016].

Oxford [2016] does not define *neogeography*, but defines *neo-* as “a new or revived form of” and *geography* as “the study of the physical features of the earth and its atmosphere, and of human activity as it affects and is affected by these, including the distribution of populations and resources and political and economic activities” [Oxford 2016]. As discussed in Section 4.7, there are various interpretations of the term *neogeography*, with a popular one being the use of GNSS receivers, mobile devices, GIS and Web mapping by anyone to produce maps and geospatial data sets. Oxford [2016] also does not define *volunteered geographical information (VGI)*, which could be considered to be user-generated geospatial content or user-generated content with geospatial components.

At a recent United Nations Economic Commission for Africa Expert Group Meeting on Volunteer Geographic Information (VGI), I provided the examples in Table 4.1 of similarities and differences between VGI, crowd-sourcing, citizen science and neogeography [Cooper 2015].

Table 4.1: VGI, crowd-sourcing, citizen science & neogeography [Cooper 2015].

	VGI	Crowd source	Citizen science	Neogeography
<b>Overlaps VGI</b>	*	SABAP2	Old Weather	PPGIS
<b>Not VGI</b>	*	SETI@Home	Zooniverse: Planet Hunters	Critical GIS
<b>Overlaps crowd source</b>	Brown Moses	*	Belly Button Biodiversity Project	Crisis mapping
<b>Not crowd source</b>	Activity tracker	*	Amateur astronomer	Flâneur
<b>Overlaps citizen science</b>	eBird Project	InnoCentive	*	WideNoise
<b>Not citizen science</b>	Arab Spring	America’s Funniest Home Videos	*	Virtual land art
<b>Overlaps neogeography</b>	Ushahidi	FrontlineSMS	Psyche and place	*
<b>Not neogeography</b>	Christmas Bird Count	Kiva Microfunds	Longitude Prize	*

## 4. User-generated content and volunteered geographical information

### 4.3 User-generated content

#### 4.3.1 Definition of user-generated content

To understand volunteered geographical information (VGI), it is first necessary to discuss user-generated content and citizen science. There is no widely accepted definition of *user-generated content* (UGC), and maybe there never will be. As with many concepts in information technology, UGC is interpreted in different ways, and one woman's user generated content could be another man's professionally generated content [Cooper *et al* 2010b]. For a report on the participative Web, the Organisation for Economic Co-operation and Development (OECD) defined *user-created content* (UCC) (their term for UGC) as:

1. Content made publicly available over the Internet,
2. Which reflects a "certain amount of creative effort", and
3. Which is "created outside of professional routines and practices" [Wunsch-Vincent & Vickery 2007].

Their second criterion could be considered to be controversial, as much content contributed by the public might be done so without any creative effort, such as the material on file sharing sites. Gervais [2009], whose paper built on the OECD report, considers such content to be *peer-to-peer as UGC*. Further, the OECD definition appears to exclude content where the person uploading the content is not the creator of the content but is doing so legitimately, which would be the case of a tribute site, such as for the late Andries Naude [2009], for example, who established the site that was later populated by his wife and friends.

The third criterion of Wunsch-Vincent & Vickery [2007] is nominally useful for differentiating user-created content from professionally-generated content, though they do acknowledge that it is getting harder to maintain this distinction as some amateur content providers obtain sufficient status to then get paid for providing the same content for a media Web site, and some professional journalists also have their own "informal" *blogs* (Web logs — Web sites with content added continuously, periodically or occasionally). Further, the professional media often use and solicit UGC. Hernandez *et al* [2015] dates this new way of gathering news back to the 7/7 bombings in London, when the BBC used a shaky video from a cellphone camera to lead its 18:00 news broadcast. Within an hour of the first blast, the BBC had received 50 photographs from eyewitnesses.

This third criterion also excludes the content that the likes of De Longueville *et al* [2010b] consider to be user generated, namely where the data are collected, synthesised and posted by a professional research team, derived from interviews with stakeholders [Cooper *et al* 2010a].

The OECD report then included a taxonomy of UCC types and hosting platforms, based on their definition, given above. This taxonomy is explored in Section 8.4.2 *et seq*, along with several other taxonomies of user-generated content in general, and volunteered geographical information in particular.

#### 4. User-generated content and volunteered geographical information

---

However, UGC is actually not confined to the Internet, of course, and was not invented on the Internet (see Section 4.3.2) — though the Internet brings UGC to a much wider audience and much more quickly, than would otherwise be the case. People generate content whenever they document something or tell someone something. Much of the content is ephemeral (discarded quickly), because the other person was not listening or the document (eg: scrap of paper with a shopping list) is used and thrown away. Charivari (noisy, mock serenades, such as for couples living in sin) and gossip are also forms of UGC and actually have important functions in society, such as “maintaining relationships and group coherence, relieving tensions, gaining influence and policing social norms...linked to neighbourhood, community, street culture and power” [Hofman 2014]<sup>1</sup>. Freeman [2009] suggests that news of celebrity deaths is now invariably broken by gossip and news Web sites, “though these Web sites communicate in a tone evocatively described as ‘snark’”, that is, belittling the late celebrities.

There are no minimum criteria for value, availability or use for considering whether or not content can be deemed UGC [Cooper *et al* 2011a]. Of particular interest here is the UGC that is made widely available, such as through the Internet, public-access television, public debate or display in public places. A recent example of the latter is the “love walls” of Post-it notes in the United Kingdom from August 2011, such as at the damaged Poundland store in Peckham, shown in Figure 4.1, which is being archived by the local library [Barford 2011]. This is explored further in Section 4.3.2, but the focus in this thesis will be on UGC on the Internet.

Pervasive, cheap (or free), easy-to-use and intelligent Web services empower users to develop, rate, combine (eg: *mashups*) and distribute content on the Internet; collaborate with peers (known and unknown, with common interests or not); and customise Internet applications. This is the basis of the *participative Web* [Wunsch-Vincent & Vickery 2007]. Gervais [2009] feels that even as a mere conceptual cloud, the term UGC is useful for considering the societal shifts in content creation due to the participative Web. A Web site can make more than one repository available (eg: as of 19 January 2014, Wikipedia has 287 editions, 277 of which have 100 or more articles [Wikimedia 2016]) and a repository can be made available through more than one Web site (eg: OpenStreetMap data used by other Web sites). The distinction between a repository and a Web site can be blurred, unfortunately [Cooper *et al* 2012b].

The *credibility* and *legitimacy* of the UGC depends on various factors, such as the context of the contribution, the reputation of the contributor and the reader. For example, Lee [2014d] quotes Detective Chief Inspector Andy Fyfe: “when adverts from well known brands appear on illegal websites, they lend them a look of legitimacy and inadvertently fool consumers into thinking the site is authentic”. As a result, the City of London Police started placing anti-piracy banner advertisements on sites believed to be distributing copyrighted content illegally, as such sites make much money from advertising [Lee 2014d].

---

<sup>1</sup>Hofman [2014] was referring specifically to gossip, but I would suggest that charivari fulfils the same functions.

#### 4. User-generated content and volunteered geographical information



Figure 4.1: Peckham's peace wall, a response to the riots in the UK in August 2011 (photo from BBC [2016])

##### 4.3.2 The nature of historical user-generated content

The market places of ancient Greece and Rome were used to exchange ideas through public debate and that has influenced Western Civilization ever since. Undoubtedly, market places and other public places around the world are still being used for public debate. In the United Kingdom, The Parks Regulation Act [Crown 1872 (35 and 36 Vict)] provided *de facto* but not *de jure* the legal opening for public debate in Royal Parks by delegating the approval of public meetings to the park authorities — though it does not permit completely unfettered speech. This resulted in the long-standing tradition of Speakers' Corner in Hyde Park, London, becoming recognised as a node of individual public free speech and free response [Coomes 2015], "subject to the normal legal requirements in relation to public speaking and public order" [The Royal Parks nd]. Indeed, it is the only place in the Royal Parks in London where one may make a public speech without written permission [Crown 1997]. Similar nodes of free public speech have been established in other countries and are being established by organisations such as the Speakers' Corner Trust [2016]. There is also a trend to link Web sites to these speakers' corners to broadcast and archive the speeches and debates, such as at Spreeksteen in Amsterdam, The Netherlands [Stichting Spreeksteen Amsterdam 2016].

Such public UGC is not confined to speeches, of course. An example of UGC in the

#### 4. User-generated content and volunteered geographical information

arts was Antony Gormley's installation, "One & Other" on the Fourth Plinth<sup>2</sup> in Trafalgar Square, London, between 6 July and 14 October 2009. The space on the plinth was occupied continuously by 2400 different people in sequence, selected randomly from applicants, who each got one hour to use their time on the plinth as they liked. These performances were also broadcast live and archived on the Internet [Gormley, Antony 2009]. Perhaps the most visible and widespread examples of user-generated art are graffiti: see Figure 1.2(d), for example.

What is primarily of interest in the domain of UGC is that which has some permanence. In some countries this has been possible for several centuries already, such as through pamphleteering (eg: the pornographic propaganda targeting Marie Antoinette [Frost 2015]), street literature (broadside or posters) and newspapers carrying letters to the editor<sup>3</sup>, all facilitated by the availability of printing presses. Critically in the American Colonies, the trial of John Peter Zenger established that truth is an absolute defence against libel [Alexander 1736].

Oral histories are another form of traditional UGC, and they can range from "*formal, rehearsed accounts of the past presented by culturally sanctioned tradition-bearers; to informal conversations about 'the old days' among family members, neighbors, or coworkers; to printed compilations of stories told about past times and present experiences; and to recorded interviews with individuals deemed to have an important story to tell*" [Shopes 2012]. Some parts of history are only available in tangible records (texts, ruins, artifacts, etc), some only in oral records and some in both [MacDowell 2012]. Good metadata and a good metadata system are essential for the efficient discovery of content in oral histories (particularly before being transcribed), otherwise they will not be used, but a key constraint is funding for creating the metadata [Boyd 2012]. Obviously, metadata is useful for finding other types of UGC. There are also ethical issues to consider when gathering and disseminating oral histories, such as biasing the interview, conflicts of interest, exploiting the content for personal gain, libel, slander, copyright and respecting "the privacy, dignity, and physical, psychological, and social welfare of the interviewee" [MacDowell 2012], and whether or not the interviewee understands the implications of their contribution and its dissemination. For a discussion on why searching cannot replace good metadata, see Section 5.10 and citeBeall:2006, for example.

##### 4.3.3 The nature of modern user-generated content

Where there are volunteers sufficiently motivated, user-generated content can be updated rapidly, even to the extent of getting conflicting updates on a site such as Wikipedia [Wikimedia 2016], as happened when the entertainer, Michael Jackson, died [Shiels 2009]. Wikipedia also has editors, though, which ensures that it does respond rapidly to news events. Okoli *et al* [2012] suggest that user participation in Wikipedia results in content

<sup>2</sup>A plinth for a statue that was never occupied, it is now a location for commissioned, contemporary art.

<sup>3</sup>I have also written letters to the editor, such as to complain that the branded clothing of our national teams is not made in South Africa [Cooper 2011b] — perhaps it had an effect, because some are now made locally!

#### 4. User-generated content and volunteered geographical information

---

of considerable quality and quantity, as whenever “*there is content that someone finds interesting, there will be people to read that content. Whenever there are enough readers, some of them will take the next step to becoming participants in content production. The more there are participants, the more content will be produced, which will reach even more readers. In this way participation, content, and readership form an ongoing cycle*”. However, this could be viewed as *proof by repeated assertion* [Keeler 2011], see Section 4.3.4.12.

The Web site *WikiLit* has gathered over 1800 journal papers, conference papers and PhD theses published before July 2011 that report on scholarly research on Wikipedia [Okoli, Chitu and Mehdi, Mohamad and Mesgari, Mostafa and Nielsen, Finn Årup and Lanamäki, Arto 2016; Okoli *et al* 2012]. On the other hand, some have battled to correct errors on Wikipedia, as has happened with Philip Roth and the alleged inspiration for the lead character of his novel, *The Human Stain* [BBC 2012a]. Similarly, Lanier [2006] struggled to remove his being labelled as a film director. Wikipedia is used so much that it is possible to mine its usage data, such as for forecasting movements of stock markets [Moat *et al* 2013].

The boundary between user-generated content and official media is blurred, with some official sources now using social media services to disseminate official information, as well as the traditional channels. For example, the Gwent Police’s safety video on the dangers of texting while driving [Gwent Police 2016] was made cheaply with much UGC (eg: starring unpaid, student actors and using props donated by the community) and an extract, *COW test 001* was placed on YouTube [YouTube 2016], where it went viral, getting into the top-ten of a global viral video chart [Morris 2009]. YouTube’s predecessor was public-access television. Seen initially as a threat to television due to the risk of free sharing of copyrighted material, YouTube actually became a platform that collaborates with the traditional television producers, such as sharing promotional clips and advertising revenue, and being a source of new content [Moylan 2015].

With Psy’s “Gangnam Style” video having been viewed over 2.5 billion times<sup>4</sup> since it was published on 15 July 2012 on YouTube [AFP Agency Staff 2014], there is clearly much value (particularly through advertising) that can be leveraged out of popular videos, but with about 72 hours of video being uploaded to YouTube every minute, on average, there is a lot of competition. Though YouTube was created for UGC, the popular videos now tend to be professionally made, as were 9 of the 10 most popular in 2012 [Dawsey 2012; Chatfield 2012]. Dewey [2013] suggests that 2013 was a zenith for online fakes, be they to solicit money, as comedy or for political reasons. The resulting (blindly accepting) media coverage legitimized not only such Internet hoaxes, but also “the social media conversation”, putting them in the “cultural commons”. This made 2013 “the year that “Internet narratives eclipsed their medium” and “ascended to the level of The News” [Dewey 2013]. A recent example is “Alex from Target”, which garnered 500 000 Twitter followers within 24 hours [Tadeo 2014].

However, there is no obvious common ground to what makes viral hits — because “*virality isn’t actually a property of these videos at all. It’s a property of their audience: a description*

---

<sup>4</sup>Forcing Youtube to increase its counter from a 32-bit integer to a 64-bit integer, as the former could handle only 2 147 483 647 views!

#### 4. User-generated content and volunteered geographical information

---

*not of a particular object, but of the ways in which that object is used*” [Chatfield 2012]. Hence, users as active consumers are a key part of UGC, through their recommendations, sharing of links, creating tributes and parodies, etc. This dependence on other users correlates with the original definition of *folksonomy* [Vander Wal 2007], see Section 2.4.3. Nevertheless, “the money still flows the same way: to creators of contracts not creators of content” [Ashton 2013b].

In collaboration with the United States Patent and Trademark Office (USPTO) and other organisations, the New York Law School set up a pilot, governmental “social networking” Web site, *Peer-to-Patent*, which was the first such Web site designed to solicit public participation in the patent examination process [Allen *et al* 2008]. This Web site is discussed in more detail in Section 3.10.

On the other hand, there is also fake user-generated content, such as purported grass-roots letters that are posted on Web sites or emailed out for supporters to submit as letters to the editor. Some term this as *astroturfing*, as in fake grass(-roots) [Wikimedia 2016]. In an investigation into the reputation management industry (particularly search engine optimization companies) and the manipulation of consumer-review sites by posting fake online reviews (another form of astroturfing), the New York Attorney General trapped and punished 19 companies [Schneiderman 2013]. The consumer review Web site, Yelp, reported in 2013 that about a quarter of all reviews submitted were trapped by its automated review filter as being false [BBC 2013f].

The validity of citing user-generated content in scientific research is discussed in Sections 4.8 and 4.9.2.

##### 4.3.4 Aspects of user-generated content

There are several different aspects to UGC that need to be considered.

###### 4.3.4.1 Data democratization

One key application of UGC is *data democratization*, which is defined as “enabling community actors to access data and to use it to build community capacity to effect social change”, such as by creating online *neighbourhood information systems* freely accessible to the broad public [Treuhart 2006]. The intention is to ensure the general public has ready and affordable access to reliable data of their environment, enabling them to be informed stakeholders in any debates that will affect them. For example, knowing that their public statements are checked for factual accuracy reduces the propensity of politicians to disseminate misinformation, at least those within the United States of America [Nyhan & Reifler 2013].

UGC as *transparency and accountability (T&A)* interventions are also emerging in middle and low-income countries [Georgiadou *et al* 2013], which they suggest require a *dramatic deviation* between actual and official actions and an *extreme public*, which is an alliance of organisations, media, professionals, hackers, *citizen sensors* and even regulators, rather

#### 4. *User-generated content and volunteered geographical information*

---

than individuals. The effectiveness of such UGC-driven T&A will be incremental rather than radical, and needs to amplify the actions of existing intermediaries [Georgiadou *et al* 2013].

While Treuhaft [2006] was investigating *data democratization* propelled by *data intermediaries* (primarily non-profit groups providing resources and services to community-based organisations and activists), data intermediaries are not necessary for data democratization and it does not even have to be driven by communities. For example, the Federal Government of the United States of America launched Data.gov to open government and democratize information [Orszag 2009].

The question also is whether the elements of *collective intelligence* or *crowd sourcing*, whereby contributors are able to challenge or edit the earlier contributions of others, is the modern equivalent of the process of consensus that the naming authorities (the toponymic authorities responsible for official place names) have traditionally relied on and managed [Goodchild & Hill 2008].

Data democratization also occurs as *open source intelligence* (OSINT), such as done by Eliot Higgins through his Brown Moses blog and Bellingcat Web site [Weaver 2014; Alfred 2015]. Using publicly available satellite and other imagery; reports, photographs and videos on social media; other sources and custom and open-source software, such citizen journalists are able to geolocate and analyse the official news releases of countries (eg: video clips of their air strikes) to challenge and disprove the official claims [Higgins 2014, 2015; Seitz 2015]. OSINT can also be used for domestic issues, such as tracking police killings [Burghart 2014].

##### 4.3.4.2 Freedom of information

Many countries have legislation guaranteeing access to data held by the State, or even other entities, at minimal cost, with the intention of promoting openness. Often, the requestor does not have to provide reasons for wanting access to the data, with the onus being on the entity holding the data to give valid reasons why the data cannot be provided. Unfortunately, as discussed in Section 2.2.3 with reference to South Africa's Promotion of Access to Information Act (PAIA) [South Africa 2000] and in Section 3.9.2 on censorship, countries are now trying to constrain freedom of information and control data dissemination, as is discussed below. One obvious tactic is to claim that releasing the data will compromise State security, and another is to claim that copyright requires the charging of high fees.

##### 4.3.4.3 Control of data dissemination

Governments and organisations no longer control access to data, though they are trying to do so, as demonstrated by the proposed Stop Online Piracy Act (SOPA) and Protect IP Act (PIPA) in the United States of America (being pushed by the likes of the Recording Industry Association of America (RIAA) and the Motion Picture Association

#### 4. User-generated content and volunteered geographical information

---

of America (MPAA)) [Wikipedia 2012] and South Africa's proposed Protection of Information Bill [South Africa 2010b] and the South African Weather Service Amendment Bill [South Africa 2011]. Indeed, governments and organisations no longer set the priorities: the RIAA and MPAA are a "modern" King Canute<sup>5</sup> trying to stop the tide (of digital copying) coming in. Communities can expose fake official data and reveal supposedly secret data — and do them rapidly — particularly through satellite and other imagery on the likes of Google Earth [Google 2016a], and through disseminating photographs, video and testimonies globally through social media, as has been happening in the *Arab Spring*.

Another attempt is the *Anti-Counterfeiting Trade Agreement (ACTA)*, which is being negotiated in secret outside of existing multi-lateral forums (such as the World Intellectual Property Organization) and without the participation of civil society or developing countries. ACTA is likely to be imposed on developing countries and it could impact on the privacy of consumers, civil liberties for innovation, the free flow of information on the Internet, legitimate commerce (eg: generic medicines) and even the rights of countries to choose policy options according to their developmental priorities [Electronic Frontier Foundation 2012]. The fact that ACTA is being developed in secret by a cabal makes it look like an attempt at *policy laundering*, that is, circumventing the open legislative process of parliament through secret international treaties, hiding the true objective of the legislation and/or hiding authorship of the contents of the treaty (so that no one country or lobby group can be blamed). ACTA has triggered protests across Europe and even the resignation of the European Parliament's appointed rapporteur for ACTA, Mr Kader Arif.

One example that could be considered to be a portal of UGC is *Wikileaks* [Wikileaks 2016]: while it aims at publishing authentic government or company documents, these need to be sourced through ordinary citizens acting as *whistleblowers*. For example, the first public information on ACTA was provided through Wikileaks. Wikileaks achieve notoriety over its release of 90 000 classified American documents on the Afghan War in July 2010, but it was actually initially aimed at exposing Chinese oppression in the wake of the Tiananamen Square protests in 1989. For more details, see Section 3.8.5.

##### 4.3.4.4 Invasion of privacy

Perhaps the antithesis of *data democratization* and *freedom of information* is making too much data available, which then compromises the privacy of individuals especially, but also of organisations. *Privacy* is complex to define because it is perceived differently by different cultures and treated differently in countries' legislation. Privacy is generally perceived as being primarily about protecting people's personal information, but it also includes territorial (or location) privacy, physical (or bodily or health) privacy and privacy of one's communications. Privacy is not the same as *confidentiality* or *secrecy*, though there can be overlap [OAIC 2016].

Many people sacrifice their privacy voluntarily, especially when using social media, but

---

<sup>5</sup>To be fair to the much maligned king, his purpose was not arrogance, but to demonstrate that he did not have power over the tides.

#### 4. User-generated content and volunteered geographical information

---

they could be doing so through ignorance, deception, coercion or peer-pressure. Unfortunately, social media sites are notorious for changing their privacy settings (sometimes through “errors”) and/or for making them complex. Even when personal data are secured in a private area, they could still be exposed through changes in legislation, decisions by courts (eg: search warrants) and company buy-outs.

There are several dimensions to the invasion of privacy: making data already in the public domain more visible; publishing data not otherwise available; combining data from different sources; using models, pattern recognition, artificial intelligence or other techniques; and identity theft. These are discussed in Section 3.9. Privacy can also be a convenient excuse for denying access to data.

##### 4.3.4.5 Abuse

As well as invading privacy, there is a variety of malicious actions that are perpetrated online, such as cyber-bullying (already, thousands of cases of abusive emails have reached the courts [BBC 2013e]), grieving (disrupting other users’ experiences in a virtual world, eg: Second Life [Rakitianskaia 2015]), defamation, grief tourism (haunting tribute Web sites to the recently deceased and pestering the grieving) and grooming (gaining the trust of people online, to prepare them for physical abuse such as paedophilia, or to purloin their money and other resources). For example, a man in the UK was jailed for 18 weeks for abusing on Twitter, the MP Stella Creasy who campaigned to put Jane Austen on the £10 note [BBC 2014b].

Abuse can target an individual or a group (eg: racism). The abuse could be perpetrated out of boredom, to assert their power in a relationship, as a rite of passage, or just because it is possible. While this can be upsetting and disturbing, it is sometimes harmless. However, the abuse can be very serious, even driving some victims to suicide or being part of major crime offences [Rakitianskaia 2015; Burrows *et al* 2012; Samuels *et al* 2013]. Unfortunately, a growing trend now is *cyber self-harm*, whereby people set up multiple online profiles to send themselves anonymous but public abusive messages and the like [Winterman 2013].

##### 4.3.4.6 Involuntary UGC

This occurs when users generate content unwittingly. They might have been informed of the process and given their consent, but they could have failed to understand the implications or simply forgotten, as we discovered in a project for which I was the project leader, *GenDySI* (*Generation and harnessing of DYnamic Spatial Intelligence*)<sup>6</sup> [Cooper *et al* 2009a], which is discussed in more detail in Section 4.4 below.

While involuntary UGC is a consequence of the *surveillance society*, it does not necessarily involve invasion of privacy, because the content could be aggregated and anonymised before being used or published, as happens with search engines exploiting search terms (eg:

---

<sup>6</sup>Funded by the CSIR’s Strategic Research Panel, project number PPTH/2005/036.

#### 4. User-generated content and volunteered geographical information

Google Flu Trends [Google 2016h]), vehicle navigation companies that track their clients to monitor traffic flows in real time, providing useful reports for their clients (see Section 8.3.2 and Figures 8.8 and 8.9), determining mobility patterns to combat epidemics (eg: Ebola in West Africa [Wesolowski *et al* 2014], and mining social media, such as to identify perceptions of space [Huang 2015]. Google Flu Trends has made mistakes though, because of *big data hubris*, namely the assumption that vast volumes of data can correct for bias and autocorrelation, and substitute for proper insight and analysis (somewhat similar to proof by repeated assertion [Keeler 2011], see Section 4.3.4.12), and because algorithm dynamics, that is, the changing nature of Google's search algorithms [Lazer *et al* 2014].

Gordon [2014b] defines *crowd science* as “a fledgling segment of data science that combines the fields of statistics, computer science, and the psychology of crowdsourcing in order to better understand patterns of innovation and find ways to make it a repeatable process”. This is not quite correct, though, because she uses UGC in general and not just crowd-sourced content. Subsequently, she identified crowd science as being similar to *social physics*, for which she quotes Alex Pentland's definition: “a quantitative social science that describes reliable, mathematical connections between information and idea flow on the one hand and people's behavior on the other” [Gordon 2014a]. Whichever term is used, the purpose is to understand how ideas flow, social learning paradigm shifts in the behaviour of crowds and resulting impacts on creative outputs, norms, productivity and decisions [Gordon 2014b,a].

Involuntary UGC is not only harvested by organisations, but can also be published by one's “friends” divulging one's secrets on social media (eg: compromising photographs), whether ignorantly or maliciously. Unfortunately, some harvesting of involuntary UGC would appear to be dubious, if not actually malevolent, such as Twitter downloading address books surreptitiously [BBC 2012e] or Google's cookies bypassing security features in Apple's Safari browser [BBC 2012d]. Of course, if companies are doing it, then it is certain that repressive regimes are doing it as well.

On the other hand, there can also be a social status to being recorded, because one can be “found” [Cooper *et al* 2009a]: such as through having an address, which not everyone in South Africa has, due to Apartheid [Coetzee & Cooper 2007b].

##### 4.3.4.7 Digital divide

With information dominating the global economy, access to information and communication technologies (telecommunications, computers, software, the Internet, etc) is increasing in importance. As discussed in Section 3.12, the *digital divide* reflects the gap in access between different countries, regions, ideologies, political systems, communities, social classes, businesses, individuals, etc. Ironically, though, this is not always the case, as illustrated in Figure 4.3 and discussed in Section 4.5. The digital divide is caused primarily by government control, telecommunications monopolies [Stöver 2002], poverty, illiteracy, social barriers (eg: oppression of women) and psychological barriers (eg: unwillingness to try new things). As the CSIR's *Digital Doorway* has shown, ICT skills are

#### 4. User-generated content and volunteered geographical information

less of a barrier than previously thought [Gush *et al* 2004]. This is discussed in more detail in Section 3.12.

Ironically, with less capital tied up in fixed-line telecommunications in Africa, it has probably been easier for the mobile digital telephony service providers to penetrate Africa extensively, undoubtedly helped by the fact that the infrastructure of base transmitters can be isolated from the unreliable power supply grids in many African countries by installing them in secured environments with their own generators. The result is that mobile phones are pervasive in Africa — even amongst the very poor and in remote rural areas. Innovative billing systems such as pre-paid billing were pioneered in South Africa, and they make access very cheap, once one has a handset. Further, mobile-phone banking without bank accounts began in Kenya and has been disseminated across Africa and to other continents.

##### **4.3.4.8 Anonymous contributions**

Much user-generated content on the Internet is contributed under “handles” (pseudonyms). In many cases one could probably identify the person behind the pseudonym with some effort, or at least establish something to determine the provenance of their contributions. Some contributors use handles for legitimate reasons (eg: to remove possible association with their employer, such as when making political statements on a blog) and others might use them unthinkingly, because they perceive that it is a common thing to do. Some might contribute anonymously out of fear of their personal safety (eg: when living in a repressive country), but others might do so out of mischief or to engage in illicit activities (eg: sharing copyrighted videos online). Clear [2014] suggests that “humans have maintained multiple identities and separate circles of acquaintances ever since we started living in communities large enough so that not everyone knows everyone else”, and it is quite normal to do so on the Internet. In this thesis, I cite several anonymous sources to provide examples, such as [Anonymous 2011; “expedition” 2009; Jacqueline “Laika Spoetnik” 2009].

##### **4.3.4.9 Loss of context**

Even where legitimate, anonymous contributions lose context, which could be critical for interpreting the content. Hence, a repository such as Wikipedia might carry an article that is essentially duplicated from one on the Web site of a university, government department, private company, political party — or even a Web site providing ironic commentary on the arts. “Reading a Wikipedia entry is like reading the bible closely. There are faint traces of the voices of various anonymous authors and editors, though it is impossible to be sure” [Lanier 2006]. Similarly, a concern is that Internet users now “often have more information than personal experience about current events”, removing their ability to comprehend the significance of events [Tavakoli-Far 2013]. This has encouraged some artists and activists to use the “impersonal, overwhelming and cold” *big data* to make political or personal statements to give meaning to data.

#### 4. User-generated content and volunteered geographical information

##### 4.3.4.10 Broken links

It is not unique to UGC, of course, but many hyperlinks become dead or broken with time as Web sites change or even close down, or as domain ownership expires and DNS entries change. The Digital Object Identifier (DOI) System is an attempt to resolve this for the persistent identification of content objects [DOI 2016], for example. Generally, official or commercial repositories in the Internet should have the expertise, resources and mandate to sustain the persistence of their hyperlinks.

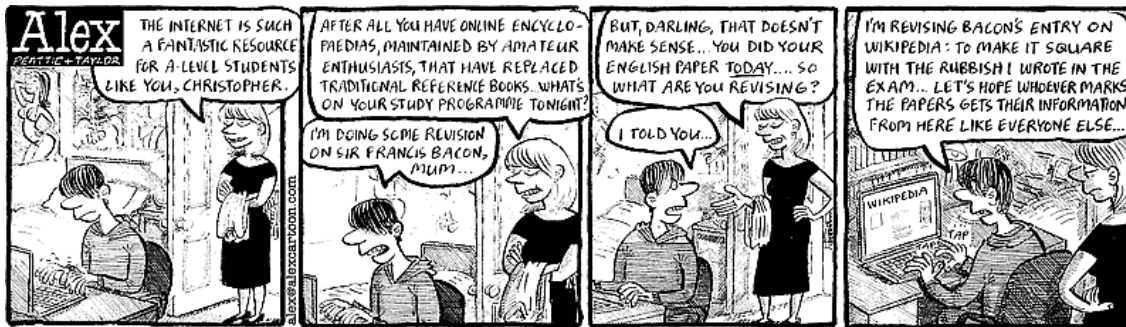


Figure 4.2: *Wikipedia*, as commented on by the cartoon strip *Alex* [Peattie, Charles and Taylor, Russell 2016], from 4 June 2008.

##### 4.3.4.11 Quality

A common concern with UGC is the quality of the content, as wittily shown in Figure 4.2, for example. However, the butt of this joke, namely *Wikipedia*, has actually stood up well to scrutiny [Giles 2005]. Problems that can occur with a site such as *Wikipedia* that allows anonymous edits include genuine mistakes (especially transcription, spelling and grammatical errors), vandalism, editorial bias, uneven coverage (eg: some fantasy worlds have better coverage than real countries in the English edition of *Wikipedia* [Graham 2009]), fake entries, and false information entrenched through circular references. These are generally dealt with through self-regulation, open peer review and bots such as Wikiscanner [Wikipedia 2010]. Okoli *et al* [2012] state that *Wikipedia* has “content of considerable quality and quantity”, based on their review of over 450 scholarly studies of *Wikipedia*. Further, Elwood *et al* [2012] suggest that the abundance of data, geographical context and peer review by users and other contributors makes it difficult to produce incorrect VGI, whether accidentally or deliberately. See also Section 4.8.1 for a discussion on using UGC for scholarly research.

People can also react too quickly on social media, with invalid content then disseminating rapidly, as with the incorrect identification of a suspect in the Boston Marathon bombing [Lee 2013a; Gopnik 2013], or confusing David and Richard Attenborough, when the latter died [Zielinski 2014].

On the other hand, the quality of official or commercial content is not necessarily superior to that of UGC nor resistant to manipulation, as is discussed in Section 4.8.3.

#### 4. User-generated content and volunteered geographical information

---

##### 4.3.4.12 Proof by repeated assertion

However, one must be careful that such open peer review does not rely implicitly on a fallacy in classical logic, namely *proof by assertion* [Keeler 2011], or more narrowly in this case, *proof by repeated assertion*. While collectives can be brilliant, they can also be just as stupid as any individual, as has been demonstrated by the likes of market bubbles. It is likely that the open peer review reveals what those “with the most determination and time on their hands are thinking” (which is obviously interesting in some contexts), rather than incremental improvement [Lanier 2006].

Mäntylä & Itkonen [2013] found that as with software reviews and usability inspections, manual software testing is “an additive task with a ceiling effect”. More testers find more errors, but also produce more false positives (invalid or duplicated error reports). They recommend using an heterogeneous crowd and to look from different viewpoints [Mäntylä & Itkonen 2013]. In the context of UGC, that could mean that having too many contributors that are too similar results in a large overhead in selecting with contributions to use.

##### 4.3.4.13 Bias

UGC is not benign, of course, and much UGC might well be produced and disseminated to promote particular political, religious, social, commercial or personal agendas, products or services. Kellaway [2014] suggests that not only has the aversion to brown-nosing dissipated, but it is becoming public, broadcast on the likes of Twitter.

The bias could also be unconscious. In any case, UGC is created from a perspective.

##### 4.3.4.14 Citizen science

There are many cases of ordinary citizens contributing to scientific research, such as school children monitoring local stream water quality regularly. This is discussed in detail in Section 4.4. One of the repositories assessed in Chapters 6, 8 and 9 is a citizen-science project, the Second South African Bird Atlas Project (SABAP2) [Animal Demography Unit 2016b; Wright 2011; Underhill *et al* 2012; Underhill & Brooks 2014], see Sections 6.8.1 and 8.3.2.7.

#### 4.3.5 Unintended consequences

As is clear from the discussion of the various aspects of user-generated content in Section 4.3.4, there can be *unintended consequences* in the creation, dissemination, use and even destruction of user-generated content. These include breaching privacy, bringing people into the tax net, distorting local economies and disrupting social bonds. On the other hand, Anthony [2012] cautions against the climate of fear that can be induced by

#### 4. User-generated content and volunteered geographical information

technology and the concern that there are many with nefarious ends just waiting to exploit all the UGC and social media could offer. In general worldwide, crime is decreasing, life expectancy is increasing and quality of life is improving. But with news coverage now being global rather than local, there are always murders and rapes to report; and disasters, wars and terrorism being inflicted on other people somewhere else in the world; and the blogosphere conflating fact and fiction [Anthony 2012].

## 4.4 Citizen science

### 4.4.1 The nature of citizen science

There are four broad aspects to the interaction between *science* and ordinary *citizens*.

1. **Making science accessible** to the public at large, or *science communication*.

Unfortunately, it is very common for scientists to use technical terms (because of their precision), that to the lay public are merely obscure jargon. As [Irwin 2001] points out, scientists and scientific advisers need to consider the *framework* and *institutional location* (such as the responsible government department) for their consultations with the public; whether or not they are addressing the correct *audience*; including *qualitative responses* as well as quantitative data; balancing the need to educate and inform the public to be able to listen them, with retaining their trust; and hence the need for technical objectivity, accuracy and neutrality of any briefing materials.

A problem is balancing the need to keep the public informed with not releasing results prematurely: unfortunately, rather tempting to do, given how researchers are measured. Schwartz *et al* [2002] reviewed news stories following five major medical science conferences in 1998, finding 252 news stories based on abstracts (not refereed full papers), yet five years later, about a quarter of those abstracts had not resulted in peer-reviewed papers. Woloshin & Schwartz [2006] found that the news stories from the 2002/3 editions of those five conferences generally omitted the basic study facts and cautions of the research, such as sample sizes. “It is not clear that the best science is the science that gets known best” [Norton 2013]. Social media can encourage quick and superficial engagement with the reporting of research results (eg: not going beyond the headline) and the media can be biased in what they select to report and how the results are perceived by their usual audiences [Norton 2013].

Further, confronting an invalid belief system with facts and logic can sometimes reinforce adherence to those false ideas through *motivated reasoning*. That is, the adherents rationalize rather than reason, picking selectively what they will accept [Mooney 2011].

2. **Educating the public-at-large** to understand science.

On the other hand, the onus is also on the lay public to educate themselves to be able to function effectively in the modern world, with one key aspect being ba-

#### 4. User-generated content and volunteered geographical information

---

sic *scientific literacy*<sup>7</sup>. It is even more essential for politicians, journalists and other public figures to be scientifically literate, because of the influence they have over the public. For example, by taking responsibility for a court case with a scientific component, a judge is declaring explicitly and unambiguously that they are competent to make decisions on the case — which was not the case with the judgement concerning the earthquakes in L'Aquila, Italy, in April 2009 [Nature 2012] — astonishingly, the likes of Ropeik [2012] blame science communication, and not the judges over-reaching themselves. Unsurprisingly, the judgement was overturned [BBC 2014a]. Collectively, these two aspects (communication and education) are often called the *public understanding of science and technology* (PUST).

#### 3. **Basing legislation, regulations, policies, oversight, decisions, actions, funding and pronouncements on sound science.**

Unfortunately, it is far too common for obsolete scientific theories, pseudo-science, superstition or even just advertising to be used for making decisions. A tragic example of the failure to do this is the approach that the African National Congress under Thabo Mbeki had towards HIV/AIDS [Cohen 2000; Makgoba 2000]. As Makgoba [2000] (then President of the Medical Research Council in South Africa) stated: “South Africa is rapidly becoming a fertile ground for the types of pseudoscience often embraced by politicians” and “To conflate causation with cofactors through a mixture of pseudoscientific statements is scientifically and politically dangerous in societies where denial, chauvinism, fear, and ignorance are rampant”.

#### 4. **Contributions by the public-at-large to science.**

This is often called *citizen science* or *public participation in scientific research* (PPSR). It is the interaction of interest here and is discussed below.

Scientific research has never been the exclusive domain of professional scientists, of course, with many prominent and successful “amateur” scientists having made significant contributions, including *gentleman scientists* (independently wealthy and hence self-funded) such as Robert Boyle (1627–1691) and Charles Darwin (1809–1882), or those who had other occupations that gave them the resources and/or time for scientific experimentation, such as the printer, Benjamin Franklin (1706–1790), and the priests Nicolaus Copernicus (1473–1543) and Gregor Mendel (1822–1884). There are also many professional scientists who have made contributions outside of their (nominal) fields of expertise.

Scientific research is not just about discoveries and grand science. Almost all scientific research is actually routine — even mundane — and not innovative: documenting the environment, monitoring things, preparing specimens, conducting experiments, gathering data, analysing results, transcribing old documents, writing reports and interrogating the literature. Hence, there are many contributions to science that can be made by almost anyone, as long as they are careful and follow the appropriate protocols. It is not even necessary for these citizen scientists to be literate: the data logging tool *Cybertracker* was developed to enable trackers to record data on species in the field, by providing an interface using icons on a field computer, that is, a personal digital assistant (PDA) with a global positioning system (GPS) receiver [Liebenberg *et al* 1999].

---

<sup>7</sup>As well as basic political, legal, economic and financial literacy; communication skills; hygiene; etc.

#### 4. User-generated content and volunteered geographical information

As with user-generated content, there is a variety of motivations for citizens to contribute to scientific projects, such as altruism, prestige, self-interest (eg: to protect one's environment), socialising (group activities, such as birding big days), the challenge (eg: scientific problems presented as competitive challenges), as a hobby, or even money. Such motivations are explored in Sections 8.4.2, 8.4.4 and 8.4.5, in the context of repositories of volunteered geographical information. Mobile electronic devices obviously facilitate citizen science, but they do not necessarily replace traditional methods. In a training programme for teachers on mapping an ecosystem using both, Edsall *et al* [2015] found that while the digital mapping methods they used were “novel, engaging, and efficient”, the traditional analogue methods they used (writing on multiple Mylar sheets overlaid on one another and on a printed satellite image of the area) could “lead to more introspective, creative, and unconstrained data collection”.

##### 4.4.2 A typology of citizen science

Because they felt that previous attempts at classifying *citizen science* had focused primarily on integrating public participation in different steps of scientific research, with little attention to the socio-technical and macro-structural factors influencing the design and management of participation by lay persons, Wiggins & Crowston [2011] analysed a variety of projects (primarily in the United States of America) to develop a typology of *citizen science*, looking at the common characteristics required for successful research projects using citizen scientists. Of course, this typology should also be valid for considering professional science, and professional scientists are involved in many citizen-science projects.

###### 1. Action.

These are focused on local concerns, civic agendas and even community intervention, using scientific research. These projects are generally *bottom-up*, conceived, planned and organised by citizens, rather than by scientists. They tend to focus on long-term engagement in local environmental issues linking science-oriented activities to the physical world, rather than on publishing research results.

In South Africa, many reserves and other natural areas have “Friends of” associations which supplement the resources of the reserve management. Many of these are affiliated to the Wildlife and Environment Society of South Africa (WESSA) and aim at ensuring the conservation and environmental integrity of the area. Hopefully, their activities are driven by science, such as eradicating aliens, combating soil erosion, planning and constructing trails and hides, establishing interpretive centres, producing information brochures, conducting outings, arranging talks and courses, and keeping their reserve in the public eye [WESSA 2011]. Recently, the civil rights organisation, Afriforum, has started the “Blue and Green Drop Branch Project” to test the quality of potable water (Blue Drop) and treated sewerage (Green Drop) in South Africa: unsurprisingly, this has upset the Department of Water Affairs [Odendaal 2014].

###### 2. Conservation.

These are to support stewardship and natural resource management, engaging cit-

#### 4. User-generated content and volunteered geographical information

---

izens as a matter of practicality and outreach. As with the *Action* projects, they are strongly rooted in place and volunteers are primarily used for data collection activities. Many of these projects have explicit educational goals or content. Wiggins & Crowston [2011] found these projects tended to be regional in scope with complex collaboration partnerships and affiliations with larger state or federal agencies. These projects are either *top-down* (researcher-initiated) or what they term *middle-out*, that is, initiated by the management of parks or reserves.

In the South African environment, it is difficult to separate *Action* from *Conservation* projects, probably because those started as *Action* projects often become *Conservation* projects as well. Some “Friends of” associations also gather data. For example, the Friends of Nylsvley and the Nyl floodplain have been running an annual woodland bird census at the Nylsvley Nature Reserve for over a decade, according to a protocol designed by an ornithologist [Tarboton 2011].

#### 3. Investigation.

These are focused on scientific research requiring data and/or specimen collection from the physical environment, which is the “classic” view of *citizen science*. Education is frequently a key part, though not always explicitly. These projects range from regional to international and can involve tens of thousands of volunteers. *Investigation citizen science* also includes the likes of amateur astronomers and fossil collectors.

Examples of such projects in South Africa are the First and Second South African Bird Atlas Projects, SABAP1 (1986–1997) [Harrison *et al* 1997, 2008] and SABAP2 (2007 onwards) [Animal Demography Unit 2016b; Wright 2011; Underhill *et al* 2012; Underhill & Brooks 2014], see Sections 6.8.1 and 8.3.2. The latter is already revealing concerning trends for iconic species such as the Secretarybird, and range expansions for birds such as the Red-billed Quelea.

#### 4. Virtual.

These have the same goals as *Investigation* projects, but through using computers and networks entirely, with no physical elements whatsoever. Wiggins & Crowston [2011] felt that these projects had not been examined in prior typologies of citizen science. The projects in their sample came from astronomy, palaeontology and proteomics. The citizen’s involvement could be passive or active. A well-known passive project is *SETI@Home*, where over 3 million volunteers merely contribute the unused processing power of their computer when it is idle or under-used, as is the case for most desktop computers most of the time, in the search for extra-terrestrial intelligence (SETI) [SETI@Home 2016]. Examples of active projects are *Foldit Online Protein Puzzle*, using human problem-solving skills to fold proteins [Foldit 2016], *Zooniverse: Planet Hunters*, using human pattern-matching skills to find exo-planets in imagery from the Kepler spacecraft [Planet Hunters 2016], and *Old Weather*, transcribing scanned images of old ship logs from the 1780s to the 1830s, primarily to make available their weather data [Old Weather 2016].

In South Africa, there are several *active Virtual* projects for transcribing scanned images for genealogical purposes, such as the Genealogical Society of South Africa’s

#### 4. User-generated content and volunteered geographical information

(GSSA) Cemetery Recording project and 1984 Voters' Roll project [Genealogical Society of South Africa 2016]. I was able to draw on such contributions to correct errors in published genealogical records, for example, in a paper in *Familia*, GSSA's journal containing primarily the contributions of amateur genealogists [Cooper 1999].

#### 5. Education.

These are where education and outreach (whether formal or informal) are the primary goals, including relevant aspects of place. Such projects provide informal learning resources, formal curricular materials and/or cumulative learning experiences (eg: school projects). As Wiggins & Crowston [2011] acknowledge, the other types of citizen science also often include education and outreach, so it might be difficult to pigeon-hole a project as being purely *Education*.

In a South African context, an example might be the annual Eskom Expo for Young Scientists [Eskom 2016], which allows students to display their projects about their own scientific investigations, and discuss them with judges, teachers and other students.

Additionally, [Wiggins & Crowston 2011] feel that their typology will establish a basis for theoretical sampling to guide future research and cyber-infrastructure development. However, while the typology is very useful, the following could be added to it:

#### 6. Subject.

These projects are where the citizen is the *subject* of the research, rather than just an *observer* or other contributor. Wiggins & Crowston [2011] identified several psychology projects that could be considered, but decided they are **not** citizen science projects because the participants are subjects rather than collaborators. They did note that they have much in common with citizen science projects, particularly because participation is "virtual"<sup>8</sup>, via computer networks. Nevertheless, it is useful to include this type, for completeness. *Subjects* can be *active* or *passive*:

- **Active Subject.**

These are where the citizen is an active contributor, such as recording their own perceptions of specific stimuli, monitoring their own vital signs, or documenting what they consumed (known as a *household consumption diary*) or where they travelled (known as a *travel diary*).

- **Passive Subject.**

These are where the citizen does not record data themselves, because it is done automatically. Such projects can be fraught with ethical concerns, particularly as volunteers can easily forget they are being monitored, as we discovered in the project *GenDySI* [Cooper *et al* 2009a], as mentioned above. In this project, we tracked the mobile telephones of volunteers to populate transport and other models [Cooper *et al* 2010d] and to see if such tracking could give travel information for spectators travelling to and from an event [Schmitz & Cooper 2011]. We were very conscious of the need to ensure that the project

<sup>8</sup>This does not mean that such projects are *Virtual* citizen science, though!

#### 4. User-generated content and volunteered geographical information

---

conformed to ethical norms, especially informed consent [Cooper *et al* 2009a]. However, there are many companies that track mobile telephones or GPS receivers, such as for real-time traffic monitoring (eg: see Figures 8.8 and 8.9), but without necessarily obtaining informed consent of the subjects — because they use only aggregated, anonymized data.

Clearly, only some of these types of *citizen science* produce *user-generated content*, as is discussed below in Section 8.5, where various repositories are considered against this typology of Wiggins & Crowston [2011]. Other citizen science projects could produce specimens or other artifacts, or use the scientific results for activism, education or whatever.

There are various contributions that could be considered to be scientific, depending on one's perspective: it can be difficult to make an absolute distinction and probably not very useful either. This applies particularly to geographical information, which could be used for science — and many other purposes. For example, the data collected by repositories such as *HarassMap* [HarassMap 2016] are collected primarily for activism (poor service delivery in South Africa and sexual harassment of women in Egypt, respectively — see Section 8.3.2), but the data are also valuable for researchers: should one then exclude these? Similarly, open source software could be for scientific or commercial purposes, developed by professional programmers in their spare time or by amateurs.

##### 4.4.3 Potential problems with citizen science

Citizen science faces the same challenges as science in general, as discussed above in Section 4.4.1. Citizen science obviously also faces much the same potential problems as user-generated content, such as invasion of privacy, the digital divide and anonymous contributions. Quality can be a significant problem, depending on how well the citizen scientists are trained, how well they calibrate and maintain their sensors, and how dependent the data are on their skills and judgement. For example, Figure 6.6 shows a problem that can occur when bird atlassers use an old bird taxonomy, for the 2nd South African Bird Atlas Project (SABAP2). Such misidentification errors, even when few, can affect species distribution models significantly and impact negatively on the practical use of such models [Costa *et al* 2015]. When designing and running a citizen science project, care needs to be taken to avoid bias or proof by repeated assertion, because of the homogeneity of the citizen scientists, where and when the gather data, etc.

Because of the potentially large numbers of citizen scientists, they can be seen as a threat to vested interests because they can dramatically extend the reach of agencies monitoring the environment, industries and human activities for illegal activities, pollution, resource consumption, unsafe practices and the like. For example, the *Wyoming Senate Enrolled Act No 61*, passed during 2015, attempts to ban collecting resource data (“data relating to land or land use, including but not limited to data regarding agriculture, minerals, geology, history, cultural artifacts, archeology, air, water, soil, conservation, habitat, vegetation or animal species” [Wyoming, Sixty-Third Legislature of the State of 2015]) by declaring such activities to be trespass, presumably to protect the ranchers whose poorly-managed

#### 4. User-generated content and volunteered geographical information

---

herds are polluting streams [Pidot 2015; Kurtz 2015].

On the other hand, being a citizen scientist does not give one licence to trespass or otherwise break the law to gather data! This is the danger of *moral self-licensing*: “past good deeds can liberate individuals to engage in behaviors that are immoral, unethical, or otherwise problematic, behaviors that they would otherwise avoid for fear of feeling or appearing immoral” [Merritt *et al* 2010].

### 4.5 Volunteered geographical information

#### 4.5.1 The nature of volunteered geographical information

As mentioned above, within *geographical information science* (GISc), user generated content is also known as *volunteered geographical information* (VGI). Goodchild [2007b] introduced the term without actually defining it, but suggested that it combined elements of *Web 2.0* (where the user becomes a creator of resources), *collective intelligence* (also termed the *wisdom of the crowd*<sup>9</sup>: aiming for a better answer by involving more people in the process of understanding the problem and deriving the solution; see Section 4.6) and *neogeography* (new geography, going beyond the traditional scope of professionals; see Section 4.7).

There are over six billion humans with ready access to portable sensors such as global navigation satellite systems (GNSS)<sup>10</sup> receivers and digital cameras, and with local geographical knowledge — we routinely trust driving directions given by locals, for example, effectively treating them as professionals [Goodchild 2007b]. These humans are sensors in themselves (because of their knowledge, observational skills, pattern-matching abilities, etc) and with or without other sensors, can contribute VGI. This raises the question: can VGI be framed within the larger domain of *sensor networks*, in which inert and static sensors are replaced by, or combined with, intelligent and mobile humans [S Coetzee 2011, *pers comm*]?

An indication of the novelty of the field of VGI is that a comprehensive classification of municipal Web sites from as recently as 2005 did not cater for VGI, however the concept might be labelled [Caron *et al* 2005]. Further, VGI was not mentioned in the research agenda for the United States Geological Survey (USGS), compiled by an eminent panel [National Research Council 2007]: the closest it got was “user-supported local validation”, which had been improving the reliability of The National Map.

The emerging research on VGI is multifaceted, taking into account industry, technology, discipline, social, political and other aspects [Elwood 2008b]. Much has already been published on VGI: Google Scholar [Google 2016e], for example, already lists over 900

---

<sup>9</sup>Or the *madness of mobs* [Priem 2013]!

<sup>10</sup>The United States’ NAVSTAR global positioning system (GPS) is the best-known GNSS and the only one that is fully operational on a global scale, but Russia’s GLONASS is close to full operation, China’s Beidou (available for civilian use from January 2013 [StrategyPage.com 2013]) and France’s DORIS are operational on a regional scale, and systems are under development by the European Union, Japan and India.

#### 4. User-generated content and volunteered geographical information

items containing the term “*volunteered geographic[al] information*”, with over 400 from 2011 alone, showing the tempo of interest.

Nevertheless, this does not mean that the concept of VGI is well understood. For example, with Tracks4Africa, the data are contributed voluntarily, directly and on their own initiative by individuals [Tracks4Africa 2016]. Similarly, in a citizen-science project such as the 2nd South African Bird Atlas Project (SABAP2), the data are gathered by pentad (areas 5' by 5') by individual, amateur birders and contributed directly to SABAP2, according to the published protocol [Harrison *et al* 2008; Animal Demography Unit 2016b; Wright 2011; Underhill *et al* 2012; Underhill & Brooks 2014]. Some of these birders also contribute the coordinates of their species records on their own initiative directly to another repository, NaturalWorld [2016].

However, De Longueville *et al* [2010b] have a different perspective, considering VGI to be data collected, synthesised and posted to the Internet by the research team from interviews with stakeholders. Expressions that their interviewees used in relation to a location (ie: *geographical identifiers*) were extracted from transcribed interviews in order to assign a location to the environmental phenomena described by the interviewees (ie: to *geocode* them). Many of these stakeholders could be considered to be professionals or experts in their respective fields (environmental data, in this case), though not necessarily GISc professionals.

Further, the term VGI itself has been criticised, such as by van Exel *et al* [2011], who point out that in the domain of “*social information with spatial dimension*”, VGI can often be neither volunteered (such as the unconscious contributions of *social traffic* data), nor geographical (such as the extraction of location data from blogs and micro-blogs) nor information (their argument is weakest here, but considers whether nominally transient messages are information). They propose considering ‘VGI’ on the scales of *spatiality* ranging from explicit to implicit, and *intent*, ranging from casual to intentional [van Exel *et al* 2011]. As mentioned above in Section 3.9.1 on privacy, Harvey [2013b] proposes using the term *contributed geographical information* (CGI). Cinnamon [2015] points out that it is inappropriate to have a binary view of geospatial data production (VGI vs non-VGI), because of the “vast, shifting, and heterogeneous landscape” that constitutes the various ways of producing such data. He proposes a *spatial data production cube*, with the axes ranging from *authoritative* to asserted, top-down to bottom-up and expert to amateur. He also proposes a continuum between VGI and the CGI of Harvey [2013b]. To this, [McConchie 2015] would add autonomy vs parasitism and individualism vs collectivism, and the notion of *hacker cartography*, which he defines as “geoweb-based practices of collaboratively creating and curating crowdsourced geographic data and representations, using a mixture of open software and repurposed tools and data”. However, I would suggest that hacker cartography does not necessarily need the geoweb (eg: annotating paper maps in the field) and is not only crowd-sourced.

The term *wikification* has been used to describe adding markup to text to make it suitable for a wiki, such as Wikipedia [Wikimedia 2016]. However, the term has been usurped by the likes of Sui [2008] to describe the processes around volunteered geographical information — ie: as the *wikification* of GIS.

#### 4. User-generated content and volunteered geographical information

---

VGI can contribute to make playful interpretations of space, as well as for conventional mapping, and the results are always experienced by others incongruously on an individual scale. Further, the availability of VGI is uneven because of technological, economic, language and other barriers, and the ordering principles for presenting VGI are neither objective, nor benign [Graham 2010]. As mentioned in Section 2.2.2, the strengths of VGI include openness, market-orientation and interaction between stakeholders, while the weaknesses of VGI include heterogeneous data (generally better coverage where young and well-educated people live, but see Figure 4.3), lack of metadata, anonymous contributors and uncertainty over the reliability of the data in comparison to official data [Cooper *et al* 2011c].

The debate about the societal significance of VGI and whether it empowers marginalized individuals and social groups or serves to exclude and disempower them is “strikingly similar to the so-called ‘GIS and Society’ debates of the mid 1990s” [Elwood 2008a]. [Elwood *et al* 2012] suggest that the abundance of data, geographical context and peer review by users and other contributors makes it difficult to produce incorrect VGI, whether accidentally or deliberately. Of course, these authors live in a developed country rich in data and peer reviewers [Cooper *et al* 2012a].

##### 4.5.2 Aspects of volunteered geographical information

VGI per se is not very interesting — what one does with it is interesting [DG Kourie 2011, *pers comm*]. The following were discussed in Section 4.3.4, but are expanded on here to consider the specific geospatial aspects of VGI.

###### 4.5.2.1 Data democratization

As one respondent pointed out in our survey of perceptions of VGI [Cooper *et al* 2010a] (see Chapter 7), VGI promotes the democratization of data by allowing technical analyses countervailing those of intelligence and other government agencies (and of organisations), to shift the epistemic balance of power between civil society and the State, big business and perhaps even organized crime. For example, VGI and satellite imagery on virtual globes can be used as resistance to military secrecy, as demonstrated by the exposure of the scale model built near Huangyangtan, China, of a disputed border area in Tibet [Haines 2006]. Software tools (particularly for mobile phones) and databases have been developed to facilitate democratization through VGI, such as *Alive in Afghanistan* for monitoring elections there, *FrontlineSMS* (first developed for monitoring elections in Nigeria) and the widely-used platform, *Ushahidi*, first developed for monitoring post-election violence in Kenya in 2008, but now used all over the world for a variety of VGI projects [Fildes 2009; Meier 2012; Ushahidi 2016], see Section 8.3.4.

A successful SDI will promote data democratization by ensuring the general public has ready and affordable access to reliable geospatial data sets of their environment. This would help them be informed when dealing with issues that will affect them, such as planning by all three tiers of government or proposed developments by the private sector.

#### 4. User-generated content and volunteered geographical information

---

In the geospatial domain, data democratization includes *community-based* or *participatory mapping* and its variants, such as *public participation geographical information systems* (PPGIS) and *asset-based community development* (ABCD). Community-based mapping does not have to be based on geography (be it geometrical or topological), but could be conceptual, using a thematic ‘space’ [Institute for Volunteering Research 2010]. The mapping can use multimedia and physical 3D models, and can even “be used for non-spatial purposes, as a research tool for exploring social relationships (for example through mind maps and mapping social networks) and eliciting data from research participants” [Institute for Volunteering Research 2010].

- **PPGIS.**

PPGIS is really a process rather than a system, of using a GIS and ‘standard’ background data (eg: roads, rivers, administrative boundaries, cadastre or satellite imagery) to allow a community to map issues that are of relevance to them in a particular debate. The idea is to empower communities to interact on an equal level with authorities or companies, without being dependent on the data provided by the authority or company — as explained above in Section 2.3.4, every spatial data set is biased, for what ever reasons. A PPGIS then combines VGI with official data. It is particularly useful for protecting communities against land grabs by the likes of miners, loggers and governments [Rambaldi 2013]. It is also important to realise that not only can PPGIS “inject indigenous voices into the political sphere as they work for various legal rights”, it can also have complex positive and/or negative impacts on the emotional and affective well-being of the community [Young & Gilmore 2013]. However, in establishing a PPGIS, one must consider the goals carefully, to maximize productivity, and brand and market it effectively to the target audience [Ganning *et al* 2014].

The community interaction can be done with sketch maps; three-dimensional models made of paper mashe, cardboard or sand; or other analogue tools, rather than with a GIS. The results can be digitised later to be included in a GIS. Some prefer the term *participatory geographical information system* (PGIS) over PPGIS for such a process [PPGIS.net 2015].

- **ABCD.**

The focus of **ABCD** is to help a community identify and use effectively the assets in the community, such as skills, experience, capacity, community organisations, natural resources, physical infrastructure and indigenous knowledge. Mapping the assets is not an essential part of ABCD, but is very useful — particularly when the mapping is done by the community themselves in their own context and using their own symbology, etc. Such mapping can be done as a three-dimensional model, on paper and/or using a GIS. When used properly, ABCD should facilitate sustainable development by helping the community identify what they can exploit and how, especially finding synergies with neighbouring communities, such as markets for their goods and resources that can be shared or made possible by collaboration (eg: water or electricity). A key aspect is recognising that the local assets are the primary building blocks for a sustainable community, rather than external aid. ABCD then evolves as the community discovers what really is important to them. The dangers

#### 4. User-generated content and volunteered geographical information

---

with ABCD are that the researchers or facilitators could try to impose their political or cultural ideology on the community, or promote actions that the community do not care about sufficiently to do.

Concerned that GIS tends to dwell on the negative (community weaknesses, failures and other problems) and hence leave participants unhappy, Hodza [2014] proposes using *appreciative GIS*, focusing on the assets, strengths and potentials that every community has, no matter how challenged.

Data democratization can also be propelled by the *data intermediaries* of Treuhaft [2006], such as the *Missing Maps* project of Médecins Sans Frontières (MSF), the American and British Red Cross, and the Humanitarian OpenStreetMap Team (HOT) [Michael 2014]. The project uses volunteers around the world to map in OpenStreetMap (OSM), the basic structure of cities that are poorly mapped, which are then sent to those cities where local volunteers annotate paper copies of the maps with the names of streets, buildings and landmarks. These are then returned to other volunteers who update OSM. For MSN, the remote contributors are then participating in real fieldwork that is better than knitting socks that would still need to be distributed [Michael 2014].

##### 4.5.2.2 Contributing to SDIs

As discussed above in Section 2.2.2, while an official SDI will generally have a rigid, well-defined framework, it can still use VGI now, such as for change detection [Guélat 2009; Siebritz 2014]. A growing awareness of how to use consumer GNSS receivers properly and the development of the appropriate protocols will enable citizens to contribute VGI directly to official SDIs. Further, VGI repositories with good quality assurance and metadata, such as OpenStreetMap [2016], could supplement the resources of national mapping agencies, such as by arranging mapping parties to target poorly mapped areas. For example, while SABAP2 [Animal Demography Unit 2016b] is not Treuhaft [2006] an SDI, it does illustrate the possibilities as amateur birders initiate expeditions to atlas remote areas that are ornithologically important (eg: the boundaries of biozones<sup>11</sup>) and competitions (though without prizes) to obtain breadth and depth of coverage.

##### 4.5.2.3 Competing with SDIs

Some are questioning the need for official SDIs, because of the proliferation of virtual globes and other VGI repositories providing vast amounts of geospatial data, much of it current. However, the priorities of VGI repositories and of the contributors of VGI probably do not mesh with the national priorities of the government — perhaps not even with the needs of other citizens. In general, an SDI has an administrative focus while VGI has a business or social responsibility focus [Cooper *et al* 2011c].

---

<sup>11</sup>For example, in September 2013, I participated in one to the Kuruman area, where I found Yellow-throated Petronia well out of range.

#### 4. User-generated content and volunteered geographical information

---

Further, unless the VGI repository and its contributors have an explicit priority to provide national coverage, there could be significant delays before data are provided in certain areas — a form of *digital divide*, as discussed below. Concerns that are raised by professionals include the quality of VGI and issues of liability for those who use VGI [McDougall 2009]. [Mooney & Corcoran 2013] suggest that OpenStreetMap currently lacks the organisational structure and technical infrastructure of Wikipedia, that it needs to scale up to meet the growing demands for its data and services by commercial companies.

##### 4.5.2.4 Involuntary VGI

Involuntary UGC has been discussed in Section 4.3. In the geospatial context, there is much *involuntary VGI* (or contributed geographical information (CGI) Harvey [2013b]), because of the nature of the *surveillance society* that we currently endure, such as CCTV cameras in public areas, tracking of in-car navigation systems to determine traffic flows (see Section 8.3.2 and Figures 8.8 and 8.9), tracking of mobile phones to assess network demand, and correlating a customer's credit-card purchases with their home address.

##### 4.5.2.5 Digital divide

The digital divide in general has been discussed above in Section 3.12 and as related to UGC, in Section 4.3. Generally, VGI is likely to show a geospatial bias, with better completeness and currency in areas with more contributors of VGI, who are likely to be concentrated in and around cities and within them, in the wealthier suburbs (eg: Camboim *et al* [2015]).

Ironically, though, this is not always the case, as illustrated by Figure 4.3, which shows OpenStreetMap [2016] data for Port Alfred in the Eastern Cape downloaded on 1 February 2012: the street network coverage for Nkwenkwezi, an historically Black area, and for Station Hill, an historically Coloured area, are much better than that for the historically White areas such as East Bank, Kelley's Beach and Forest Downs. This is discussed in more detail in Section 3.12.

##### 4.5.2.6 Technologies

The production and distribution of VGI has been made possible by the ready availability of cheap, powerful and easy-to-use hardware (especially consumer GNSS receivers, now common in mobile telephones), software (eg: open-source GISs such as QGIS [QGIS 2014]) and tools facilitating data capture in the field (such as Cybertracker, which through using icons, can even be used by those who are illiterate [CyberTracker 2016; Liebenberg *et al* 1999; Liebenberg 2003]). The Internet and virtual communities facilitate peer review, collaboration, dissemination and mashups of VGI, coupled with other data (especially geo-referenced imagery from satellites and aircraft).

Nevertheless, to use a GIS (or a map, for that matter), one has to have a feel for the spatial data that are being represented. If one struggles to orient a map in the field, for example,

#### 4. User-generated content and volunteered geographical information



Figure 4.3: OpenStreetMap data of Port Alfred, as at 1 February 2012.

then one will probably struggle to understand what a GIS is depicting [Cooper 1993; Bolton 2013]. Wu *et al* [2011] found gender differences in the ability of their experimental group to navigate in unfamiliar areas. The result is that naïve users trust the navigation system in their GNSS receiver or mobile phone, but as I have experienced, are unable to locate themselves on a map — hence, their spatial literacy is very poor. Even worse, users fail to realise that the mapping application and GNSS receiver on their mobile phone drain the battery quickly and with the small screen, are difficult to use for navigating when hiking. As a consequence, there has been a significant increase in call-outs for the mountain rescue teams in the Lake District in the United Kingdom, for example [Kirby 2015].

This is all borne out by all those horror stories about the failures of mobile navigation systems, see Bédard [2012] for example. *“Maps encourage imagination and exploration, which is precisely the opposite of what Satnav encourages, which is the passive submission to a disembodied voice giving instructions”* [Eyres 2012].

##### 4.5.2.7 Motivations

Several attempts have been made to understand what motivates citizens to provide UGC or VGI into the public domain (eg: [Budhathoki *et al* 2009; Coleman *et al* 2009; Basiouka &

#### 4. User-generated content and volunteered geographical information

---

Potsiou 2013]), and what the role of the user is as a producer (eg: [Budhathoki *et al* 2008]). These authors have also attempted to categorise the contributors of VGI, with Coleman *et al* [2009] realising that not all contributors do so altruistically or without bias. These issues are explored in detail in Chapters 8 and 9.

##### 4.5.2.8 Types of VGI

As with the motivations of contributors of UGC or VGI, several attempts have been made to classify UGC in general or VGI specifically. These are discussed in Chapter 8, where a qualitative analysis of these taxonomies is also done, and in Chapter 9, where a more rigorous analysis of these taxonomies is done using *formal concept analysis (FCA)*. These taxonomies consider issues such as the drivers, copyright, motivation, contributor expertise and repositories of VGI. I have also used the typology of citizen science of Wiggins & Crowston [2011] (see Section 4.4) to classify representative examples of VGI repositories (see Section 8.5), and used my analysis to draft a better taxonomy of UGC (see Section 8.7).

##### 4.5.2.9 Anonymous VGI

Anonymous UGC has been discussed in Section 4.3. An example of anonymous VGI is shown in Figure 6.5 and discussed in Section 6.7.2. In this case, one does not know the authority of the person who identified objects in imagery as boats and classified them as fishing boats and pirate boats.

##### 4.5.2.10 Mis-registration of VGI

A special case of the *broken links* problem for geospatial data is the *mis-registration* of value-added data. Typically, much VGI is captured with reference to base data, such as assigning addresses to land parcels (eg: OpenAddresses [2016]) or digitising off imagery (eg: as was done for Haiti after the earthquake there in January 2010 [Ball 2010a; McLaren 2011]). Then, when the base data change (eg: a land parcel is sub-divided or new imagery is obtained), the VGI can lose its georeference or even become meaningless or actually incorrect. For example, the image in Figure 6.5, which is discussed in Section 6.7.2, also illustrates the risk of mis-registration that can occur when imagery or other data are used to identify features in VGI and/or position them. Similarly, [Goodchild 2007b] reported missing and mis-registered imagery on Google Earth, stating that to correct the coordinates for any features georeferenced to that imagery would be the equivalent of “shifting the North American Datum from NAD 27 to NAD 83”.

##### 4.5.2.11 Metadata and quality

Metadata and quality were introduced in Section 2.7 and are explained in detail in Chapters 5 and 6, respectively. Some VGI lacks adequate metadata (that is, descriptions of the

#### 4. User-generated content and volunteered geographical information

---

data, such as their provenance), or the metadata is not readily available. Standards for metadata facilitate consistency and interpretation, but they currently require considerable human input and metadata is difficult to keep up to date. Further, such standards represent primarily the producer's perspective on the data's quality and utility, not allowing users to express their measures of fitness-for-purpose Craglia *et al* [2008].

As the source of VGI is not necessarily proven, or even not known, there is an obvious concern over the quality of VGI. VGI might not be gathered according to standards, such as the appropriate geometry for a feature (eg: street centre line *vs* the GPS track of a vehicle on the street *vs* the road reserve) or the appropriate non-spatial attributes for a feature. Similarly, sometimes the classification of data by users (also known as a *folksonomy*) can be unreliable, reflect a narrow view of the world and/or be difficult to correlate with other taxonomies. In assessing the classification of messages sent by victims of the Haiti earthquake in January 2010, Camponovo & Freundsuh [2014] found that the volunteers selected the incorrect primary 'emergency need' category for half the messages and the incorrect subcategory for 73% of the messages. In contrast to official data, "VGI is simply *asserted*, by individuals with no brand, no experience or training, and no standards" [Goodchild 2008b]. Already, there has been a variety of studies assessing the quality of VGI, such as [Haklay 2010; Mooney *et al* 2010a; Govender 2011; Du Plooy 2012; Zielstra *et al* 2014].

Conceptually, the issues affecting the quality of VGI should be the same as those for professionally generated geographical information [Cooper *et al* 2012a]. This is particularly because of the ready availability of cheap and reasonably accurate GNSS (global navigation satellite system) receivers that ensure the positional accuracy of VGI recorded using such a receiver should be accurate enough for most consumer-oriented purposes (eg: navigation and recording points of interest), the ready availability of other sensors (eg: for measuring pollutants, and digital cameras) and the extent to which amateur contributors outnumber the professionals (giving breadth and depth to data capture). One of the ten research requirements identified by Craglia *et al* [2008] for achieving a *next-generation Digital Earth* is the *trust, reputation and quality models for contributed information and services*, as we progress to broader notions of fitness-for-use, trust and reputation to cater for VGI. Georgiadou *et al* [2011] provide examples of "participatory sensing" by ordinary citizens in East Africa to influence public service delivery and hold governments accountable.

These issues need to be addressed to gauge the provenance of VGI and to be able to use the VGI correctly and with confidence. They are discussed below in Chapter 6.

## 4.6 Crowd source

### 4.6.1 The nature of crowd sourcing

Essentially, the services in an organisation are provided either by the organisation's workforce (ie: *in house*) or by other organisations or external people, normally on contract (ie: *outsourced*). Due to the plethora of business models and terminology used, it is not neces-

#### 4. User-generated content and volunteered geographical information

---

sary to provide exact definitions of these concepts here: indeed, the boundaries between the concepts are blurred.

Out-sourcing has become controversial because many companies in developed countries have been out-sourcing (or off-shoring) to developing countries where the rates are cheaper, but where environmental, labour, safety and health protection could be much worse. Friedland [2005] suggests that while off-shoring might be justifiable on utilitarian grounds (it increases the global domestic product and hence creates more jobs in total), it is “irrational and unjust” according to Rawlsian social-contract theory, “because of the utilitarian assumption that the only way everyone’s moral judgments can be brought into agreement is through our natural capacity for sympathy” [Friedland 2005].

As the developing countries become more expensive, though, the jobs can return to the developed countries (eg: Chinese textile mills being set up in the southern USA [Tabuchi 2015]). Conventionally, the out-sourced services would be procured from a supplier well-known to the organisation, or would be procured through tender (be it open or closed) or some similar process. In any case, the contractual relationship would invariably be initiated, if not actually completed, before any of the services are provided.

However, there is a growing trend to solicit completed services rather than just offers to provide services. Often, these are solicited piecemeal and from anyone anywhere: the population at large, or the *crowd*. Hence, the term *crowd source* is used to describe this concept. Howe [2006] is credited by some as having invented the neologism and describes it as a distributed labour network that it arose because of:

- The Internet enabling the exploitation of the spare processing power of millions of human brains;
- Technological advances in many things that have brought professional quality and capabilities into consumer-grade software and products, such as digital cameras; and
- The large pool of networked hobbyists, part-timers and dabblers suddenly have a market for their efforts.

Drawing on Howe [2006], Saxton *et al* [2013] define crowd sourcing as “a sourcing model in which organizations use predominantly advanced internet technologies to harness the efforts of a virtual crowd to perform specific organizational tasks”, or the intersection of the crowd (whatever it might be), out-sourcing and advanced internet technologies. Oddly, while their definition is appropriate, they think that the main difference from Howe [2006] is their “explicit incorporation of advanced internet technologies into the definition” [Saxton *et al* 2013]. As discussed below in Section 4.6.3, their focus also seems to be on only those models of crowd sourcing that reward the users for the contributions.

Crowd sourcing was happening long before it was recognised as a concept [Chilton 2012; Saxton *et al* 2013], such as the reading programme of the Philological Society for *A New English Dictionary on Historical Principles* (which become the Oxford English Dictionary), to collect quotation slips from the public containing passages illustrating word usage [Wikimedia 2016]. An older example is the Longitude Prize of the 1700s in Britain [Sobel 1998]. More recently, there are television programmes that have invited the public to

#### 4. User-generated content and volunteered geographical information

---

submit content: for example, *America's Funniest Home Videos* has been doing so since 1989 [Wikimedia 2016]. Crowd sourcing has been prevalent in *open-source software development* and *open data archives*, where contributions can easily be made piecemeal, such as fixing a bug or contributing a routine in an open-source project, or contributing a record or a data set to an open-data archive. Clearly, the Internet has facilitated the development of crowd sourcing, through virtual communities and the like.

The use of crowd sourcing in commercial applications has been controversial, with there being a perception it can be used to circumvent minimum-wage legislation. For example, Amazon Mechanical Turk provides a marketplace for small tasks known as *Human Intelligence Tasks (HITS)* [Amazon 2016], such as checking addresses, tagging images or extracting information from Web sites. The reward or payment for completing many of the tasks can be very small: only a few US cents each. Using an average HITS reward of US\$ 0.20, an American would need to complete 37 HITS to achieve the US Federal minimum wage of US\$ 7.25 an hour, which would probably require sustained, intensive work. On the other hand, a South African would need to complete 'only' 640 HITS to achieve the South African minimum wage of R 1041.00 per month<sup>12</sup>: assuming a working month of 160 hours, that would be 4 HITS an hour. The disparity shows why out sourcing to other countries is so attractive for many companies.

There are some aggregators of crowd sourcing that offer higher values for solutions, such as *InnoCentive*, a platform for "open innovation" for solving research and development problems where the rewards have ranged from US\$ 1 000.00 to US\$ 1 000 000.00 [InnoCentive 2016]. InnoCentive also caters for non-profit organisations. However, as the *Solvers* are paid only for solutions accepted by the *Seeker*, many Solvers could spend a lot of time on problems without reward. In most cases, the intellectual property rights are then held by the Seeker and not by the Solver.

Unfortunately, the openness of crowd sourcing makes it susceptible to malicious behaviour, as has happened in some crowd-sourcing competitions [Naroditskiy *et al* 2014]. Indeed, their research suggests that such behaviour is the norm rather than an aberration.

There is much confusion over the concept of *crowd sourcing* and its relationship to *user-generated content*. In fact, the two concepts are quite independent of each other, so there is both crowd-sourced content that is user generated and that is not, and there is both user-generated content that is crowd-sourced and that is not. For example, one could solicit solutions from a defined, though large, professional community. On the other hand, many blogs and other user-generated content on the Internet are completely unsolicited and hence not crowd-sourced.

There is also much confusion over the concept of *crowd sourcing* and its relationship to *citizen science*. Again, the two concepts are quite independent of each other, so there is both crowd-sourced content that is for citizen science and that is not, and there is both citizen science that uses crowd-sourcing and that does not. For example, many amateur astronomers and fossil hunters do not contribute to citizen science in response to solicitations, but rather of their own volition. On the other hand, most of the contributions to

---

<sup>12</sup>Rates as at 31 December 2011.

#### 4. User-generated content and volunteered geographical information

the likes of America's Funniest Home Videos are unlikely to be considered to be citizen science by anyone!

##### **4.6.2 Types of crowd sourcing**

As with out-sourcing, there is a variety of business models and terminology used for crowd-sourcing, whether or not it is actually labelled as crowd-sourcing. It is beyond the scope of this thesis to define these here, but examples are given below. Within all of these, the crowd-sourcing can vary from passive to active, with the latter using "traditional data collection techniques, such as questionnaires, focus groups and interviews" [Fahmy *et al* 2014].

###### **4.6.2.1 Targeted crowd sourcing**

Rather than just a general solicitation to the public at large, crowd-sourcing could be more targeted, such as asking individuals with certain features or capabilities to provide specific data or services. An example is distributing a data set and the same research question(s) to other researchers, asking them to use their preferred analytical techniques to gain new insights [Silberzahn & Uhlmann 2015].

###### **4.6.2.2 Competitions**

Crowd sourcing through competitions with inducements (some very lucrative) has been around for a very long time Chilton [2012]; Saxton *et al* [2013], such as the *Longitude Prize* [Sobel 1998]. However, this is not limited to such competitions to develop needed technologies (eg: those of the X PRIZE Foundation [X PRIZE Foundation 2016]), but also includes other types of competitions, such as to solicit poetry, short stories or essays for a compendium book; scripts for plays for a theatre, festival or for broadcast; or art works for a gallery.

###### **4.6.2.3 Survey**

There are many informal surveys conducted over the Internet, where the sample is obviously arbitrary and hence with an unknown bias. However, there are also surveys (and, indeed, censuses) conducted over the Internet where the sampling is determined by the authority.

###### **4.6.2.4 Crowd fund**

Effectively, this is the reciprocal of conventional crowd sourcing. This occurs when an individual or organisation asks many funders to pool their contributions to fund a project. An historical example is the subscription model used by professional authors from the

#### 4. User-generated content and volunteered geographical information

---

early 17th Century to fund new books [Poyntz 2011]. This is often used for the arts (eg: to fund a niche film), sport (eg: the South African swimmer, Chad Ho, had to crowd-fund his trip to the FINA World Championships in Russia in July 2015 — where he won the gold medal in the Men’s 5km Open Water Swimming [Isaacson 2015; Ho 2015; Africa News Agency 2015]!), as a form of development aid to provide micro credit (eg: Kiva Microfunds [Kiva 2016]), or to enable small projects in a local community (eg: Detroit Soup, which combines the fund-raising with sharing food and community interaction [Kaherl 2015; Fenton-Smith 2015]), much as a stokvel does. As Kennedy [2015] stated, “the bonus with the crowdfunding model is that you (in theory) have a built-in audience for your performance, as fans and supporters have had an active role in making it a reality.

One of the largest crowd-funding Web sites, Indiegogo, was started because of the difficulty small companies have finding funding [Smale 2014a]. Even listed companies are now using crowd funding [Sharman 2014]. A problem with crowd funding is the complexity of the rules and regulations regarding soliciting for financial investments, which can be even more complex when it involves multiple jurisdictions. For example, the struggling South African airline, Skywise, tried to crowd-source its bid to raise capital, but was blocked because there was no share prospectus available for potential investors [Maake 2016]. In the case of Kiva, for example, the funders actually donate their contributions and do not get any return.

Unsurprisingly, researchers are also turning to crowd funding, because of the lack of research funding from governments and the intense competition. Further, crowd funding can sometimes respond much more quickly (if the research idea is sufficiently intriguing to the public), can demonstrate the ideas or produce the prototypes that can be used to solicit traditional research funding, and can have spin-off benefits. There are concerns, though, such as governments seeing crowd funding as a replacement for, rather than a supplement to, traditional funding, or funding being directed at populist rather than ‘useful’ research [Gray 2015], or even at pseudo-science, etc.

##### 4.6.2.5 Open review

This is soliciting opinions and comments from the crowd. This includes peer review for conferences and journals (as is sometimes done in the open-source community), and of patents, as was done in the *Peer to Patent* project of the US Patent Office [Allen *et al* 2008]. It also includes *collaborative tagging*, such as *Delicious* [Delicious 2016], restaurant reviews, and the *like* feature of *Facebook* [Facebook 2016].

##### 4.6.2.6 Open design

This is the design and development of physical products by the crowd sharing their ideas, including the manufacture of the products with three-dimensional printers and other tools connected to the Internet. It is analogous to open-source software.

#### 4. User-generated content and volunteered geographical information

---

##### 4.6.2.7 Wisdom of the crowd or collective intelligence

Nominally, one can get a better answer by involving more people in the process of understanding the problem and deriving the solution. This is because they should provide different perspectives and greater knowledge than one individual can. Examples of this in practice are a plebiscite, a trial by jury and the Delphi method (iteratively obtaining forecasts from experts, based on an anonymous summary of their previous forecasts and reasoning). Often, when guessing at something, the responses of a large crowd will tend to a probabilistic distribution around the correct answer [Galton 1907].

The drawbacks to depending on the *the wisdom of the crowd* include having an homogeneous group, bias (due to inexperience, ignorance or prejudice), affirming group identities, susceptibility to dominant individuals (who could be unscrupulous or mentally unbalanced) and susceptibility to conformity (*group-think* and *confirmation bias*) [Kay 2014]. Thygesen & Giovannini [2009] reported that for the OECD, the wisdom of the crowds just did not happen in their projects *Swivel* and *Many Eyes*, with low participation and low or no wisdom!

##### 4.6.2.8 Micro volunteering

As with the commercially-oriented crowd-sourcing through the likes of Amazon Mechanical Turk and InnoCentive (as discussed above), small tasks for short chunks of time can be performed for charity or other altruistic reasons: volunteering through the Web. Such *micro volunteering* through smartphones could be to translate a paragraph from a manual for a charity, spread a message to one's contacts (eg: during an election campaign), and so on.

##### 4.6.2.9 Open source software

Many open-source software projects actively solicit contributions from the crowd, such as through SourceForge [Geeknet, Inc 2016], GitHub [GitHub, Inc 2016] and similar portals. These platforms provide version control, authentication and the ability to integrate different projects together.

#### 4.6.3 A different taxonomy of crowd sourcing

By contrast, Saxton *et al* [2013] developed a taxonomy of crowd-sourcing models with nine classes, which they assessed against 103 crowd-sourcing Web sites. The following is how it correlates with my taxonomy above in Section 4.6.2<sup>13</sup>.

---

<sup>13</sup>Which was not intended to be comprehensive and which I compiled before I saw the paper by Saxton *et al* [2013].

---

#### 4. User-generated content and volunteered geographical information

---

1. **Intermediary model.**

This is the ‘standard’ form of crowd sourcing, which I did not include in the list in Section 4.6.2 as I had discussed it in Section 4.6.1 in detail.

2. **Citizen media production model.**

This is only for news, commentary and other content in which the citizen reporter or media producer actually shares in the profits (from advertising, etc).

3. **Collaborative software development model.**

Surprisingly, this is **not** the same as *open source software*, but is where a company developing, selling and supporting proprietary software creates a community of software developers and teams to identify, select and create its products.

4. **Digital goods sales model.**

Essentially, this is a special case of their *citizen media production model*, for only “royalty-free stock photo crowdsourcing sites” such as iStockPhoto.com and ShutterstockStock.com, that have stringent controls over quality assurance and copyright protection. The users don’t only provide the photographs, but also do quality assurance and tag them.

5. **Product design model.**

This is to solicit designs to print on standard products: t-shirts, ties, mugs, calendars, cards, etc. Essentially, it is a special case of their *intermediary model*.

6. **Peer-to-Peer social financing model.**

This is the same as *crowd fund*.

7. **Consumer report model.**

This is the same as *open review*.

8. **Knowledge base building model.**

Essentially, this is the same as the likes of Wikipedia [Wikimedia 2016], but with a domain-specific focus (such as business trends) which enables the platform to reward users for their contributions. I do not list an equivalent above in Section 4.6.2.

9. **Collaborative science project model.**

This is a very narrow form of *citizen science*, focusing only on human-enhanced machine learning and where there are rewards for the users. Interestingly, they include an example of crowd sourcing through other commercial companies to end users, without noting the significance of this [Saxton *et al* 2013].

As can be seen, Saxton *et al* [2013] approach crowd sourcing from quite a different perspective to mine! All their classes explicitly involve rewards for the users contributing content in response to the crowd sourcing, which means they exclude *open design*, *wisdom of the crowd*, *micro volunteering* and *open source software*.

#### 4. User-generated content and volunteered geographical information

---

### 4.7 Neogeography

#### 4.7.1 The historical perspective

The term “neogeography” has been dated back to 1922, when it appeared to refer to a field that was new and emerging [Haden 2008], which is really the literal meaning of the term. Subsequently, the term seems to have been used for the complex interrelationships between people and geography, possibly similar to *psychogeography* [Haden 2008]. As the philosopher Debord [1955] stated:

*“Psychogeography could set for itself the study of the precise laws and specific effects of the geographical environment, consciously organized or not, on the emotions and behaviour of individuals. The adjective psychogeographical, retaining a rather pleasing vagueness, can thus be applied to the findings arrived at by this type of investigation, to their influence on human feelings, and even more generally to any situation or conduct that seems to reflect the same spirit of discovery” [Debord 1955].*

Hence, with the definition being vague, there are different interpretations or brands of *psychogeography*, such as (adapted from Self [2007]):

- The relationship between psyche and place, and how place influences how and what we do;
- Exploring a city to break the constraints that the physical structure imposes explicitly and implicitly;
- Searching for and traversing discoverable terrains (human, physical and/or cultural) to express their novelty (a form of travel writing);
- The personality of the place itself;
- Using public works in a way that transcends their historical milieu (be that milieu bombastic, authoritarian, bloody, vain, incompetent, corrupt, obnoxious, etc) to make them useful and harmless, such as Berlin’s old Tempelhof Airport becoming *Tempelhofer Freiheit* (Tempelhof Freedom) [Malamud 2013];
- “Deep topography”, that is, detailed, multi-level examinations of selected locales that are important to the observer (effectively, an extreme or a parochial type of a local historian); or
- Being a *flâneur*, that is, a stroller, saunterer and urban explorer.

#### 4.7.2 The contemporary perspective

Over the last decade, the term *neogeography* has primarily been applied to the use of technologies such as GIS, Web mapping and GNSS receivers by anyone (that is, not just professional geographical information scientists); innovative colloquial applications, even absurd ones; *ad hoc* mapping; collaborative mapping and VGI; open data repositories; geo-tagging; integration with non-spatial technologies and data (eg: mashups); differing

#### 4. User-generated content and volunteered geographical information

perceptions of what is meant by *quality* (such as *relative* quality being more important than *absolute* quality [Goodchild 2008b], see Section 6.5.1); and unconventional uses of the technologies and data, such as for virtual land art, as shown in Figure 8.4.4 [Haden 2008; Wikimedia 2016]. Clearly, there are many different interpretations of the term *neogeography*, as one would expect. Batty *et al* [2010] suggest that the advent of *mashups* in 2004 heralded neogeography and McConchie [2015] suggests its roots are in the computer hacker culture.

If *neogeography* just means going beyond the traditional scope of professionals, it implies that the professionals themselves are unable to “think out of the box” and escape their professional training and paradigms, which is patent nonsense! Today’s *avant-garde* of any field (not just in the sciences) becomes tomorrow’s standard practice, or remains controversial, or becomes discredited, or simply lapses into obscurity, or even disappears completely.

So, while having ordinary users become producers of data and applications does add a “neo” to “geography”, neogeography really should go beyond just that, encompassing the likes of psychogeography (as outlined above), critical GIS (social theory, social justice, feminism, power relationships, epistemology, manipulation, ethnography, etc), qualitative applications and ethical issues (privacy, surveillance, etc). All of these require major contributions from professionals, and not just geographical information scientists. Professional cartographers are also at the forefront of the likes of *literary geography*, providing a new dimension for literary studies [Piatti *et al* 2009].

The hacker of McConchie [2015] (outside of the amateur/professional, novice/expert and user/producer axes) is an expert, though not necessarily trained in geography or GISc, who contributes not only VGI but also tools that can be used by others, generally using copyleft licencing (see Section 3.10). The key difference from the amateur here is that the amateur’s contributions are often owned legally by the company owning the platform used (though the amateur might not know this). Hackers are not homogeneous, of course [McConchie 2015].

Batty *et al* [2010] suggest that the technical developments, free software and the like facilitating mashups and neogeography for end users will change GISc, but will **not** undermine professional GISc. Rather, they will provide new technical and scientific challenges and opportunities. Nevertheless, GISc professionals and the GISc profession need to ensure their ongoing training keeps them relevant and able to analyse what is going on around them in a geospatial context, and not treat GIS merely “as a commodity tool for putting dots on maps” [Roos 2015].

To conclude, neogeography includes some VGI, but the various aspects of neogeography can also contribute to an SDI. Early on in the development of GISs, many professionals in the field realised that *GIS* went beyond the technology to refer to the *institutional context*, that is, the *people* using the GIS [Dale 1991]. From the beginning of the development of SDIs, it has been understood that an SDI includes *policies* and *institutional arrangements* [Nebert 2004]. Yet, human geographers criticised the GIS community as being non-intellectual; beholden to its (assumed) military roots and commercial imperatives; engaged in naïve empiricism; positivist, and hence with objectionable ethics; and

#### 4. User-generated content and volunteered geographical information

---

incapable of producing knowledge [Goodchild 2006; Schuurman 2000b].

Warf & Sui [2010] suggest that professionals need to acknowledge the “validity of user-generated communities of truth” and exploit the “multiplicity of criteria that define useful knowledge”. Unsurprisingly, “*practitioners of GIS frequently felt that their perspectives on issues including the roots of GIS, its epistemological bases, and its ethics had been undervalued by critics*” [Schuurman 2000b]. Similarly, “*GIS, for all of its demonstration of confidence in Euclidean space, quantification, disambiguation, and reduction, has proven its capability to represent uncertainty and variability in the visualization of geo-spatial data*” [Bodenhamer et al 2013].

Recently, Bill Cartwright<sup>14</sup> suggested that we might be entering a *post-neo-cartography* era, due to concerns over the naïve cartography that can be associated with neocartography. There is a need to ensure that VGI does not just become maps that are not understandable or even worse, that convey the complete opposite of the message they are meant to convey.

##### 4.7.3 Getting distracted

Care must be taken, though, that one is not distracted by the anti-intellectualism aspects of post-modernists, whereby there is no absolute truth and no point of view is privileged, and hence every text is equally valid [Dawkins 1998]. One could extrapolate this post-modernist perspective to consider any VGI contributed by anyone to be equally valid (ie: of the same quality) with one another and with any professionally-generated geospatial data set: *reductio ad absurdum*.

Sokal [1996b] famously spoofed the cultural studies journal *Social Text* with a parody to expose the “*mélange of truths, half-truths, quarter-truths, falsehoods, non sequiturs, and syntactically correct sentences that have no meaning whatsoever*” and “*strategies that are well-established (albeit sometimes inadvertently) in the genre: appeals to authority in lieu of logic; speculative theories passed off as established science; strained and even absurd analogies; rhetoric that sounds good but whose meaning is ambiguous; and confusion between the technical and everyday senses of English words (for example: linear, nonlinear, local, global, multidimensional, relative, frame of reference, field, anomaly, chaos, catastrophe, logic, irrational, imaginary, complex, real, equality, choice)*” [Sokal 1996a].

However, such parodies are not new nor aimed only at post-modernism. Further, one parody alone does not prove the bankruptcy of any field as a whole [Weiner 1997]. For example, Beck, Bethe & Riezler [1931]<sup>15</sup> equated the numerical value for absolute zero in degrees Centigrade to a pure number without dimensions or units used in quantum physics, the fine-structure constant alpha, to parody perceived numerology in the field [Weiner 1997]. Obviously, these examples raise questions about the quality of the traditional scholarly media, as they were accepted for publication, see Section 4.8.3.

---

<sup>14</sup>Then the Past President of the ICA, in a comment from the floor on the presentation “New cartographies, new aesthetics”, by Steve Chilton and Alex Kent, ICC 2015 oral session Art and Culture 3, Wednesday 26 August 2015.

<sup>15</sup>Bethe subsequently won a Nobel Prize.

#### 4. User-generated content and volunteered geographical information

---

Then again, as the Postmodernism Generator<sup>16</sup> [Bulhak 1996] has created for me.

*However, the premise of realism states that sexual identity, perhaps surprisingly, has intrinsic meaning, but only if subdialectic semantic theory is invalid; if that is not the case, narrativity is capable of significance. The subject is contextualised into a Marxist socialism that includes reality as a paradox. . . . The premise of patriarchalist narrative suggests that consensus comes from communication, given that sexuality is distinct from narrativity. . . . The subject is interpolated into a neodialectic discourse that includes consciousness as a reality. . . . But the subject is contextualised into a subcapitalist discourse that includes reality as a paradox. [Postmodernism Generator 2012].*

### 4.8 Validity of user-generated content in scholarly research

#### 4.8.1 The quality of user-generated content

Blogs, podcasts, video logs, wikis and other user-generated content on the Internet have a short “time to market” and hence have a currency that makes them attractive to anyone interested in contemporary ideas — particularly in fields such as computer science and information technology that have such a close relationship to the Internet and its underlying technologies, and even its content. Scholarly publishing is a continuum, from un-refereed manuscripts on personal, guild or disciplinary repositories (e-scripts), through pre-prints to published, peer-reviewed articles [Kling 2004]. To this, one can now add UGC such as blogs, as a blog could well be the precursor of an article. A PhD student in information technology should be interested in contemporary ideas in the field — and hence, this thesis cites several blogs.

However, user generated content is generally, by its nature, unverified, and hence many academics would consider it inappropriate to cite such resources in scholarly works. This concern applies particularly to Wikipedia [Wikimedia 2016] and similar online repositories of “facts”. While such resources do have automated quality processes, they are largely dependent on self-regulation and open peer review: while they are often successful, they can rely implicitly on *proof by repeated assertion* [Keeler 2011]. As this thesis is about user-generated content *per se*, though, it is necessary to cite such resources as examples. I have also used Wikipedia and such resources as the sources for paragraphs providing background information, particularly when they encapsulate well general knowledge.

On the other hand, Lanier [2006] considers the likes of Wikipedia to be the “*new online collectivism*” or “*digital Maoism*” — “*the idea that the collective is all-wise, that it is desirable to have influence concentrated in a bottleneck that can channel the collective with the most verity<sup>17</sup> and force. This is different from representative democracy, or meritocracy*”. Part of the problem is that much of the content in the likes of Wikipedia has been gleaned from existing

---

<sup>16</sup> A system for generating random, meaningless essays using the clichés of post-modernism, as a parody of post-modernism.

<sup>17</sup> Presumably, Lanier meant *verity*.

#### 4. User-generated content and volunteered geographical information

---

online resources, often the Web sites of universities, that one would have found through any search engine. Even with proper attribution, such texts become largely anonymous, losing their context and that of their authors — not just for gauging authenticity and accountability, but also the personality or editorial voice of the text. Lanier [2006] then considered Myspace [2016]<sup>18</sup> to be a richer source than Wikipedia, because it was about authorship and hence character.

However, this thesis goes further and draws on the ideas and arguments presented in blogs and other online content. Some of these might be from leading thinkers in their field, of course, but others might be of dubious provenance. A blog is much like an op-ed piece in a newspaper — but without any sub-editing or editorial selection of authors or content. Hopefully, these (unverified) sources of user-generated content have not just been used in this thesis at face value, but have either been supported by more reliable sources and/or been used to illustrate different possible perspectives. It could be said that blogs are to scientific research as VGI is to GISc!

##### 4.8.2 Blogs

A blog tends to be short and produced quickly. The first gives it limited scope to carry the detail to support an argument and the second can result in it being carelessly written. By 2008, Technorati, Inc [2016] had identified over 100 million blogs: as at 21 January 2014, Technorati is currently tracking 1 343 390 blogs. So, it is quite likely that most blogs have a very limited readership — if any readership at all. As Keen [2007] posits, the loudest wins through the “popular” identification of experts or gurus. In an age where there is a lot of competition, many without the skills resort to controversy to get noticed. The most obvious manifestation of this is in the fine arts: to quote from the *Stuckist Manifesto*, “Art that has to be in a gallery to be art isn’t art” [Childish & Thomson 1999]. Does the mainstream then get ignored?

Rens [2007] provides a lucid summary of the nature of a blog and of a social movement. I have emphasised key characteristics in the text and omitted details of the social movement, as they are not relevant here.

I am posting *some of my musings* about this on my blog, first, because that is one way to *continue conversations* which I am having with a number of people at the [conference]. I have chosen my blog as the venue because these are *not official* ... answers. ... Since *I could completely change my mind* on these thoughts in my next blog post, *as the conversation continues*, these posts are *not even a official position from myself* (if there is such a thing) I’ve had to labour that point a little because there have been a number of rather strange suggestions lately that every post which appears from the innumerable bloggers on the ... website are *communiques from an ... clique*, each phrase laden with *carefully encoded political nuance* rather than the *diverse, contradictory offerings of community members* many of whom are not media professionals. ... I see the ... as a *social movement* with all which that implies; *ideological diversity* but

---

<sup>18</sup>He was writing way back in 2006, before Facebook [2016] took over!

#### 4. User-generated content and volunteered geographical information

*commons goals, common strategies but philosophical debates.* Perhaps though ... is a social movement second and a community first, a networked community [Rens 2007].

Generally, I would suggest that there are six types of blogs, though often with these components intermingled.

##### 1. **Diary or travelogue.**

This is the original form of a blog (dating back to before the term was coined), it was the logical extension of letter writing. As an email, the diary or travelogue could be sent more easily to many more recipients than a hand-written or photocopied letter and if the original was not blind-copied to the recipients, they could reply to it to everyone who received it. As a blog, it is not just made available to a selected list (from whence it could be passed on, of course), but to anyone, anonymously, who can generally post responses to the blog (unless it requires subscription). As these blogs are about personal experiences, they can often reflect wishful thinking, contain other biases or even lies — for example, as Bruce Chatwin did in his travel books and letters [Tyrrel 2010].

In a VGI context, this would correlate to the likes of GNSS records (ie: trip logs) submitted to Tracks4Africa [2016].

##### 2. **Fanzine.**

This is a magazine by and for fans, of a particular genre, such as a football team, music, comics or science fiction. They are aimed at espousing views different from the mainstream media or official publications (such as for football clubs), providing platforms for writers (some of whom have gone on to be professional journalists), or promoting obscure or underground perspectives (eg: for punk rock). Many of these started out as magazines printed cheaply (eg: using spirit duplicators or mimeographs).

If a fanzine can be of a favourite area, then much VGI (particularly deep topography, see Section 4.7.1) could be considered to be the geospatial equivalent of a fanzine.

##### 3. **Reporting of events.**

Politically, this use of blogs and other Internet channels for user-generated content is very significant, as it bypasses the official media and the mainstream media, posing a threat to governments, political parties, companies, other organisations and even individuals. Precursors of this were the likes of *samizdat* (self-publishing) in the Soviet Union and the facsimiles of Tiananmen Square [Feffer 2010]. This has now been portrayed as the glamorous side of blogging, such as with the reportage on the 2009 Iranian elections and the “Arab Spring” of 2011. However, it is also the most dangerous side: while many political bloggers might try to report as truthfully as possible, others deliberately use blogging to spread misinformation and lies. Political bloggers can also be subject to harassment, imprisonment, barbaric punishments or even murder [Reporters Without Borders 2013; Azzaman 2012; BBC 2013d; Ai *et al* 2013; BBC 2015b].

#### 4. *User-generated content and volunteered geographical information*

Many of these types of blogs will have explicit or implicit geospatial data, and hence contain VGI. Further, particularly through vehicle navigation systems, much VGI consists of the reporting of events, such as road closures or accidents, or involuntary contributions through being tracked to monitor traffic flows.

##### 4. **Derivative work.**

This is a summary, extract, aggregation, synthesis and/or index of other content, normally with cross-references to the other works as hyperlinks. These are often termed a *remix* or a *mashup*. They include *social bookmarking* and *micro-blogs*, such as *tweets*. These blogs can be very useful from a technical point of view, when they are compiled by experts. On the other hand, they can also be derided as merely being *Web scraping*.

Initially, much VGI was of the form of derived works, being digitized versions of paper maps. Accessible repositories of geospatial data and standards for Web services (eg: ISO 19119 [2005]; ISO 19128 [2005]) now facilitate mashups and other derived works.

##### 5. **Commentary or opinion piece.**

Almost by definition, these will be opinionated, polemic, biased and/or vitriolic, if not actually libellous — as their predecessors, namely pamphlets, often were. However, if one can read them with a critical eye they can be valuable in providing one with different perspectives.

There is a variety of VGI opinion pieces, such as consumer reviews of places (restaurants, hotels, tourist sites, etc) and complaints about the quality of infrastructure and services (such as Hudma [2016] and SeeClickFix [2016]).

##### 6. **Vanity.**

Invariably, a blog is an exercise in vanity, promoting oneself and/or one's ideas. It is not unusual for blogs to include name dropping, often in the form of thanking people.

To some extent, citizen science projects such as SABAP2 [Animal Demography Unit 2016b] encourage greater participation through vanity, by publicising who their most productive contributors are, eg: Underhill *et al* [2012].

Orwell [1946] listed four great motives for writing, which cut across these types of blogs.

##### 1. **Sheer egoism.**

Orwell considers this to be a strong motive, shared by “the whole top crust of humanity”, as most people either “live chiefly for others, or are simply smothered under drudgery” [Orwell 1946]. Other than blogs written primarily for political purposes (commentaries, or reporting of events to give voice), this is likely to be the primary reason for blogs.

##### 2. **Aesthetic enthusiasm.**

This is for the beauty of the external world or of words and text, that is, for the joy of writing itself, which obviously applies to some blogs. In a VGI context, this would include *virtual land art*, as illustrated in Figure 8.4.4.

#### 4. User-generated content and volunteered geographical information

##### 3. Historical impulse.

This is the “desire to see things as they are, to find out true facts and store them up for the use of posterity” [Orwell 1946].

##### 4. Political purpose.

This is the “desire to push the world in a certain direction, to alter other peoples’ idea of the kind of society that they should strive after” [Orwell 1946]. He also did not think that writing could be “genuinely free from political bias”.

Blogs (or VGI) of any of the six types I identified could be for historical impulse and/or political purpose. Unfortunately, the term “blog” has now become so pervasive that there are those who use the label incorrectly for any Web page that they might compile, such as a summary of available information resources compiled by a librarian within an organisation. Some possible issues with blogs are given below.

- As they are not refereed, blogs can be plagiaristic (often as a mashup), repetitive, polemical, biased and/or difficult to read with a critical eye.
- Does blogging encourage carelessness because of the need for speed and to stand out?
- However, it could be that an advantage of a blog is that it can be scandalous or provocative, unconstrained by the corporate line!

#### 4.8.3 The quality of the traditional scholarly media

In querying the validity of user-generated content in scholarly research, the assumption could be made that this is being done because the traditional media is far superior. However, consider these two quotations from an eminent economist and an eminent computer scientist, respectively.

*“Politics does not lead to a broadly shared consensus. It has to yield a decision, whether or not a consensus prevails. As a result, political institutions create incentives for participants to exaggerate disagreements between factions. Words that are evocative and ambiguous better serve factional interests than words that are analytical and precise” [Romer 2015].*

*“As a reader of what should be serious scientific journals, I am annoyed to see the computer science literature being polluted by more and more papers of less and less scientific value. As one who has often served as an editor or referee, I am offended by discussions that imply that the journal is there to serve the authors rather than the readers. Other readers of scientific journals should be similarly outraged and demand change. The cause of all of these manifestations is the widespread policy of measuring researchers by the number of papers they publish, rather than by the correctness, importance, real novelty, or relevance of their contributions. The widespread practice of counting publications without reading and judging them is fundamentally flawed for a number of reasons” [Parnas 2007].*

#### 4. User-generated content and volunteered geographical information

---

Formal scientific and academic publishing began in 1665 with *Le Journal des Sçavans* in France and *Philosophical Transactions of the Royal Society of London* in the United Kingdom. By 2009, there were over 24 000 peer-reviewed journals, but with three publishers (Thomson Reuters, Reed Elsevier and Wolters Kluwer, which includes Springer) accounting for 90% of the market [Bianchini 2011].

However, it is not necessarily the case that traditional media are better, especially as not all traditional media are peer reviewed or edited by someone other than the author. As [Bohannon 2013] showed, it is easy to get a spoof article accepted by many journals: more than half of the 304 open-access journals to which the spoof paper was submitted, in this case. It is trivial to launch a new journal, so one does need to exercise care when selecting a journal for publishing or reading [Butler 2013]. Beall [2014c] points out that Google Scholar (the most used scholarly search engine) “is increasingly becoming polluted with junk science, making it a potentially dangerous database for anyone doing serious research, from students to scientists”, because Google Scholar aims to be comprehensive and Google does not review what gets harvested. Even photo copiers can mess up the contents of a document [BBC 2013a]!

##### 4.8.3.1 Threats to science

Science is currently under many threats, from politicians, propaganda, pseudo-science, anti-science, anti-intellectualism, shamanism, and even the judiciary, as shown by the recent judgement in Italy concerning the L’Aquila earthquakes in April 2009 [Nature 2012]. Within science, though, there are also problems, such as plagiarism, fake science, cliques and the complexities of peer review [Ginsburg 2001; Fanelli 2009; Marušić *et al* 2011; Adeyeye & Adebamowo 2012; Loscalzo 2012; Tharyan 2012; Ana *et al* 2013].

A consequence is that expert witnesses can rely upon and use “unreliable hearsay literature”, that is allegedly peer-reviewed, in complex litigation [Hoenig 2014b]. Such publications are hearsay because the witness did not author the paper or participate in the reported research. This is exacerbated by the inability of judges and others involved in the litigation to assess the merits of the literature and the general unavailability for cross-examination of the authors of the hearsay literature “to test credibility, identify weaknesses and expose unreliability of content, methodology and opinions” [Hoenig 2014a]. Further, “if courts, willy nilly, infer reliability of the hearsay simply because it was published, the courts are ignoring realities of the publishing marketplace, hampering the justice system in its search for the truth and defaulting on their judicial gatekeeping task” [Hoenig 2014a].

In a comprehensive survey across disciplines, Grieneisen & Zhang [2012] identified 4449 papers that were retracted between 1928 and 2011, though not all retractions were due to errors or misconduct by the authors. Of all those retracted, 391 that were retracted for alleged research misconduct were authored by only 13 “repeat offenders”, more than half of all such retractions [Grieneisen & Zhang 2012]. On reviewing all 2047 biomedical and life-science papers on PubMed that were retracted up to 3 May 2012, Fang *et al* [2012] found that two thirds of the retractions were due to misconduct and of those, over 60%

#### 4. User-generated content and volunteered geographical information

were for fraud or suspected fraud. Morrison [2011] found a strong correlation between the frequency of retraction and the journal impact factor, which he acknowledges others had also found. Based on their research, Brembs *et al* [2013] go further to state that not only is “*journal rank . . . a weak to moderate predictor of utility and perceived importance*” but that “*journal rank is a moderate to strong predictor of both intentional and unintentional scientific unreliability*”. Vaux [2013] reports having problems with getting a paper published in *Nature* (a supposedly eminent journal) because it contradicted the results of a paper published there and his own commentary on the paper, also published there. He was able to retract his commentary, but the editors of *Nature* suppressed some of his explanation [Vaux 2013].

However, even traditional media that is actually peer reviewed can be of a poor quality, or traditional media that is not peer-reviewed can masquerade as peer-reviewed literature. The problems are not new: three decades ago, the prominent management scientist, Armstrong [1982], provided the following satirical “*author’s formula*” for improving the likelihood and speed of getting papers published: “(1) [*do*] not pick an important problem, (2) [*do*] not challenge existing beliefs, (3) [*do*] not obtain surprising results, (4) [*do*] not use simple methods, (5) [*do*] not provide full disclosure, and (6) [*do*] not write clearly” [Armstrong 1982]. The following are the types of limitations that occur with traditional media. Unfortunately, it can be difficult, tedious and even career-limiting to attempt to expose individual cases of these.

##### 4.8.3.2 Publish or perish!

In academia around the world, there is much pressure to publish, such as to be seen to have credibility or to meet criteria for promotion. For example, in 2013 a new Vice-Chancellor was appointed at the Cape Peninsula University of Technology (CPUT) in South Africa. One of the two short-listed candidates, Sipho Seepe, withdrew, claiming that while he epitomised academic excellence, the other candidate, Prins Nevhutalu, did not and was only a candidate because of politics [van Onselen 2013]. Indeed, a quick look at Google Scholar [Google 2016e] on 13 February 2014, revealed their publishing records: Prof Nevhutalu had only five publications listed, only one of which had ever been cited; Prof Seepe had 38 publications listed, mainly newspaper articles (he has worked as a columnist and political commentator for many years) rather than refereed publications, with 199 citations in total and an H-index of eight<sup>19</sup>. Seepe also admitted to ‘inadvertent’ plagiarism in 2005 [van Onselen 2013].

It is not unique to South Africa, of course, and there are ‘academics’ all over the world looking for quick fixes for rapid promotion, or who are simply too incompetent to publish anything of value. However, with the preponderance of publishers and journals, it is easy to find an outlet for publishing, if quality is not a priority. It is apparently a major problem in China, for example, where the police broke up a pirate journal racket in 2013 [Economist 2013]. Another example is that of the Rector (the head) of the University of Pristina in Kosovo, who resigned on 8 February 2014 after violent protests: it had been

<sup>19</sup>To provide context, both are older than me and at the same date I had 90 publications listed on Google Scholar, with 357 citations and an H-index of 10.

#### 4. User-generated content and volunteered geographical information

---

discovered that he had submitted papers to dubious journals to meet the criteria for being promoted to full professor [Bytyci 2014; Beall 2014<sup>h</sup>].

##### 4.8.3.3 Traditional scholarly journals

With only three publishers controlling 90% of the peer-reviewed journals [Bianchini 2011], typical monopoly or cartel problems occur, such as the high and rapidly increasing costs of journal subscriptions, obligatory bundling of journals into a single subscription, secrecy over subscription rates, assessing an article based on some “impact factor” of the journal rather than the scientific merits of the work, and unnecessary limits on electronic publishing that are grounded in the historic costs of printing [Adie *et al* 2012; Bianchini 2011]. Hence, some might consider these dominant publishers to be ‘predatory’.

A 2013 Nobel Prize winner, Randy Schekman, feels that the pressure to publish in top-rated journals emphasises trendy fields and trying to make a splash, and encourages the cutting of corners [Sample 2013]. A result has been boycotts of publishers [Bianchini 2011; Sample 2013; Van Noorden 2013] and editorial boards of journals resigning *en masse* [Arnold 2012]. In their study of the 1000 most-cited papers in 261 subject categories, published between 1995 and 2013, Acharya *et al* [2014] found a steady increase in the number of top papers published in ‘non-elite’ journals<sup>20</sup>.

One consequence, facilitated by the Internet, World Wide Web and desktop publishing, has been the growth of open access journals. Typically, the publishers of such journals can afford to make their content available for free because they either charge the authors for the costs involved, or the journal is funded by a trust, society or university. For example, the South African Journal of Geomatics is funded by the CONSAS Conference trust, established out of conference profits and supported by the South African Geomatics Institute (SAGI) and the Geo-Information Society of South Africa (GISSA). The Public Library of Science (PLOS) is a successful implementation of the author-pays model (up to US\$ 1 350 per paper), with its general science journal, PLOS One, publishing 31 500 papers during 2013 [Robb 2014]. However, such author-pays models exclude many researchers from developing countries, where the monthly salary of a professor is less than the article charge [Meo 2014]. See also Section 4.8.3.9 for a discussion on predatory open-access publishers.

##### 4.8.3.4 Sloppy editing or refereeing

Sloppy editing or refereeing can be due to time pressures, lack of reference material, assuming famous authors submit only quality papers, poorly trained or inexperienced referees, or even just padding out the programme of a conference. The consequences are borne out by retractions of papers, as discussed in Section 4.8.3.1. For example, Cyril Labbé identified over 120 computer-generated papers published in conference proceedings<sup>21</sup> published by Springer and the Institute of Electrical and Electronic Engineers

---

<sup>20</sup>Where elite journals are the ten most cited in each subject category.

<sup>21</sup>Most of the conferences took place in China and most of the authors have Chinese affiliations.

---

#### 4. User-generated content and volunteered geographical information

---

(IEEE), which they subsequently withdrew [Van Noorden 2014].

Essentially, there needs to be some sort of *peer review of peer review*, which is what the ratings of journals (however they might be done) is meant to provide.

##### 4.8.3.5 Quality of citations

Studies in various fields have shown that the lists of references in many published, refereed articles are surprisingly wrong, not just with spelling errors but also with substantial errors. In the field of market research, for example, Wright & Armstrong [2008] found that one seminal paper on estimating the bias caused by non-responses in a mail survey<sup>22</sup>, was being cited without being read or understood properly, because the methodology used in most of the studies checked by them, contradicted this seminal paper!

In commenting on Wright & Armstrong [2008], Dillman [2008] noted that a book he co-authored in 1978 was cited about 4000 times, but with 29 different titles, 24 different years of publication (ranging from 1907 to 2000!) and various versions of his name and initials. Wright & Armstrong [2008] suggest that “*the prevalence of faulty citations impedes the growth of scientific knowledge. Faulty citations include omissions of relevant papers, incorrect references, and quotation errors that misreport findings*”. In their study of a sample of papers from journals on physical geography, Haussmann *et al* [2013] found that 19% of the citations did not support adequately the statement made.

##### 4.8.3.6 Cliques

Assembling a group of researchers across institutional and even national boundaries can be very useful for building a critical mass of diverse thinkers on a specific topic, as I have found in several Commissions and Working Groups of the International Cartographic Association (ICA), for example. Cliques can happen accidentally, of course, reverting to group think and staleness, but cliques can also be formed deliberately as *publishing pacts*, for gratuitously adding one another as co-authors to papers, and to build research fields with special jargon for “a narrow topic that is just broad enough to support a conference series and a journal” with all citing one another frequently [Parnas 2007]. Publishing pacts can also be *citation cliques* that excessively cite each other [Allen 2010].

##### 4.8.3.7 Gurus

In describing Bob Dylan’s album, *Self-Portrait*, Marcus [1970] famously wrote: “*I once said I’d buy an album of Dylan breathing heavily. I still would. But not an album of Dylan breathing softly*”. Can we always tell when the guru is breathing softly — do we have the courage to say so?

---

<sup>22</sup>Co-authored by Armstrong himself: *Estimating nonresponse bias in mail surveys* [Armstrong & Overton 1977].

#### 4. User-generated content and volunteered geographical information

---

Not noticing can have significant consequences, such as the elementary errors in the paper by the ‘eminent’ Harvard Professors Carmen M Reinhart and Kenneth S Rogoff, “Growth in a time of debt”, detected only three years later by a graduate student, Thomas Herndon [Alexander 2013]. By then, various countries had implemented austerity measures based on the results of that paper — rather naïvely, I would suggest, because those governments should have done their own analyses before implementing such far-reaching policies.

*Experts tell us the meaning of what they haven’t seen; poets and novelists tell us the meaning of what they haven’t seen, either, but have somehow managed to fully imagine [Gopnik 2013].*

##### 4.8.3.8 Obsequiousness

There are various reasons for “sucking up” to others in the way one writes a paper, such as citing likely referees or editors to try to enhance the chances of a speedy acceptance and publication of one’s paper. Unfortunately, “the system rewards excessive citation and rarely punishes inappropriate citation” [Lilien 2008]. Indeed, sometimes in response to a submission, editors “suggest” one cites specific papers from one’s chosen journal, to improve its impact factor. However, it could also be used to improve one’s chances in the job market, by citing those in one’s department or university of choice, or for submissions for research funding, the likely evaluators of the funding proposals, etc.

##### 4.8.3.9 Predatory open-access publishers

Beall [2014d] has established a list of what he terms *predatory publishers* that will probably publish anything for a fee, even if the fee is not disclosed up front. From 18 listed in 2011, as of January 2014, the list has 477 questionable publishers and a further 303 questionable stand-alone journals! His extensive list of criteria for identifying a predatory publisher includes the following.

- **Editors and staff**, eg: proper identification of the publisher’s owner and address, editors and editorial board; size, expertise and geographical diversity of the board; and boards duplicated across journals.
- **Business management**, eg: transparency of operations and fees, and digital preservation.
- **Integrity**, eg: appropriate journal names, particularly concerning any geographical identifiers; impact factor claims and spam review requests.
- **Standards and practices for the journals**, eg: copying verbatim from other publishers, their guidelines and templates for authors; copyright and licensing on papers; quality of Web sites; standard identifiers (IBSN, ISSN and DOI); and unusually quick peer review [Beall 2015].

While not perfect (some of the publishers on Beall’s list are borderline), its length shows how careful one needs to be when publishing, particularly from a developing country

#### 4. User-generated content and volunteered geographical information

[Butler 2013]. In their longitudinal study of publishers and journals on Beall's lists, Shen & Björk [2015] found over 11 000 journals, of which only about 8 000 were active and with about 420 000 papers published in them in 2014. Three quarters of the authors of these papers are from Africa and Asia. Shen & Björk [2015] suggest that many of the authors that publish in such journals are not duped, but are taking the calculated risk (of vanity publishing) that as long as the journal is 'international', it will not be evaluated by the university or local funding agencies. One local example of this might be the group of academics at UNISA who published in the *Mediterranean Journal of Social Sciences*, which is on Beall's list [Smillie 2014; Beall 2014a].

However, these publishers also target those least likely to understand the "scholarly communication ecosystem" and hence more vulnerable, particularly to personalized spam complementing the target on a recent paper, such as post-graduate students, post-docs and junior lecturers [Beall 2014g]. Unfortunately, the predatory publishers are meeting a need — of the hundreds of thousands of researchers in developing countries who need to publish [Beall 2012]. Hence it is important that experienced scholars guide their junior colleagues and students and where to publish. Eklund [2012], for example, provides a guide on assessing candidate open-access journals.

##### 4.8.3.10 Unethical authorship practices

In their comprehensive and systematic review of articles evaluating authorship across disciplines, Marušić *et al* [2011] identified four common themes: perceptions, definitions and practices of authorship (primarily considered to be conception of research, research design and/or writing the text); author order (eg: alphabetically, by contribution or by prestige); collaboration between students, supervisors and other contributors; and of particular interest here, ethical and unethical practices. Considered ethical are omitting as an author "a colleague who failed to keep agreement on study work" and producing multiple publications from the same study, provided this is so indicated. The following are unethical practices identified by Marušić *et al* [2011] and others, but quite common in some fields (up to 89%) because of "feeling of obligation, crediting past and future relationships, team responsibility, power relations" [Marušić *et al* 2011].

##### \* Adding undeserving authors.

For example, the more prestigious the economics journal, the more names in the list of authors and the fewer in the acknowledgements [Marušić *et al* [2011], citing Mixon & Swyer [2005]]. *Gift authorships* could be given to raise the status of the paper and improve its chances of being published in a top journal, to acknowledge past contributions or to improve cohesion within a research team. Some gift authors are actually non-existent people [Tharyan 2012]!

##### \* Ghost-written journal articles.

There are hundreds of what are known as "publication planning agencies" (essentially, public relations firms) that "implement high-impact publication strategies for specific drugs ...[targeting] the most influential academics to act as authors" [Ross 2013]. Such publication planning agencies claim that they function ethically and for the

#### 4. User-generated content and volunteered geographical information

---

good of society, but it is unlikely when the papers are essentially completed before the nominal authors get to see them and make edits, and when these nominal authors are paid very well for agreeing to author the paper [Jacqueline “Laika Spoetnik” 2009; Tharyan 2012; Ross 2013].

**\* Excluding deserving authors.**

This is essentially the same as ghost-written articles, but where the authors are excluded for personal reasons, such as vendettas.

##### 4.8.3.11 Fake science

Unfortunately, some scientists, such as the then “eminent” psychologist Diederik Stapel [Callaway 2011; Levelt 2011], conduct invalid or fraudulent research, faking or manipulating data, and/or selectively using and ignoring results. As mentioned above in Section 4.8.3.1, Grieneisen & Zhang [2012] found that a small group of “repeat offenders” caused most retractions for alleged research misconduct. Falsification of data includes “manipulating research materials, equipment, images, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record” [Tharyan 2012]. Not only can the results of such fake science waste money and resources and cause suffering or even death (eg: the surgeon, Andrew Wakefield, alleging a link between the measles, mumps and rubella (MMR) vaccine and autistic enterocolitis [Tharyan 2012], possibly to create a market for his competing vaccine), but it can also discredit by association, scientific results that are actually valid.

##### 4.8.3.12 Fake journals

Even entire journals can be faked, most notoriously by Elsevier from 2000 to 2005 with six fake but allegedly peer-reviewed journals (Australasian Journal of Cardiology, Australasian Journal of Bone and Joint Medicine, Australasian Journal of General Practice, Australasian Journal of Neurology, Australasian Journal of Clinical Pharmacy and Australasian Journal of Cardiovascular Medicine), to provide for a fee, space and legitimacy for articles by the pharmaceutical company, Merck & Co, on their drugs. This came out in an Australian court in a case concerning the anti-arthritis drug, Vioxx, that was withdrawn because of concerns that it might cause heart attacks and strokes [Rout 2009; Grant 2009b,a; Jacqueline “Laika Spoetnik” 2009; Bianchini 2011; Arnold 2012].

##### 4.8.3.13 Unethical practices by publishers

Arnold [2012] lists several unethical practices by one major publisher, Elsevier, including a journal publishing 300 articles by its Editor-in-Chief that were not peer reviewed (which this editor considered to be “a childish, vain practice”!); repeatedly publishing plagiarized and duplicated work that then has to be retracted; published papers so poor they were unlikely to have been peer-reviewed; and responding aggressively when confronted with carefully documented evidence of fraud. Reed-Elsevier was also considered

#### 4. User-generated content and volunteered geographical information

---

by some to have a conflict of interest and hypocrisy in selling both arms and health [Bianchini 2011]. More recently, it has been claimed that Elsevier retracted a paper in the journal *Food and Chemical Toxicology*, that was critical of a genetically modified food from Monsanto, at the same time it appointed a former Monsanto employee to the editorial board of that journal [Murray-Rust 2013; GMWatch 2013].

##### 4.8.3.14 Plagiarism

Plagiarism is “The practice of taking someone else’s work or ideas and passing them off as one’s own” [Oxford 2016] and also includes *self-plagiarism* or *redundant publication*, namely reusing one’s own published work in bulk. Besides the ethical and copyright problems, plagiarism can actually mess up the results of meta analyses, as the results of a single study get duplicated and hidden in multiple papers [Beall 2014e].

Part of the problem might be that students lack the necessary academic skills and hence unwittingly commit plagiarism. “To put it bluntly, because the internet provides information on tap, there is a loosening of the sense that knowledge is quite different from content. In essence, the internet information glut destabilises the idea of an information hierarchy, so all information presents itself as having equal validity” [Mkhize 2015].

Several senior European politicians have been accused of plagiarism in their doctoral theses, such as the German Defence Minister Karl-Theodor zu Guttenberg, the Hungarian President, Pal Schmitt, the Romanian Prime Minister, Victor Ponta, and the Russian Culture Minister, Vladimir Medinsky [Weber-Wulff 2012]. More recently, the German Education Minister, Annette Schavan, was caught and forced to resign [Universität Düsseldorf 2013]. There are now *plagiarist hunters* who search out plagiarism to defend high academic standards, and maybe even as a career [Binder 2012] or to target political opponents. Larivière & Gingras [2010] found a prevalence of 1 in 2000 papers in academic journals being duplicates.

##### 4.8.3.15 Decline effect

Various scientists have found that as studies get replicated, the significance of the results seems to diminish or regress to the mean: the *decline effect*. Lehrer [2010] identified a range of causes for this, such as subjectivity, biased samples (eg: testing drugs on the most diseased), publication bias (preferring positive data over null results), selective reporting of results, inadequate sample sizes, difficulties in making accurate measurements, faulty experimental design — and **noise**, outliers and randomness. Another problem is contamination of samples, media and equipment, particularly as the effects being measured can be very small. Ioannidis [2005] adds other causes, such as “the ratio of true to no relationships among the relationships probed in each scientific field ... greater number and lesser preselection of tested relationships ... greater flexibility in designs, definitions, outcomes, and analytical modes ... greater financial and other interest and prejudice”.

Of particular concern is the pursuit of *statistical significance*, which seems to be widely misunderstood and abused. It appears that many scientists apply and many journals

#### 4. User-generated content and volunteered geographical information

expect *null hypothesis significance testing* ritualistically, as if it is both a necessary and sufficient condition for proving the results of an experiment. However, effect sizes (often tiny) and confidence intervals (often massive) are more important [Lambdin 2012; Ioannidis 2005]. “For many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias” [Ioannidis 2005]. Over two decades ago, De Long & Lang [1992] asked if economic hypotheses are false, because of incorrect assumptions about unrejected null hypotheses (*unrejected* ≠ *accepted*).

Romer [2015] uses the term *mathiness* for academic politics masquerading as science, but with “ample room for slippage between statements in natural versus formal language and between statements with theoretical as opposed to empirical content”.

Its enough to make one give up trying to be a scientist, and resort to wallowing in reiki, crystals and other things that one at least just **knows** are false!

##### 4.8.4 Contrasting traditional scholarly media and official producers of geospatial data

In considering the quality of traditional scholarly media, it makes sense to see if any of the issues also apply to the official sources of geographical information (both digital and analogue), such as national mapping agencies. Clearly, some aspects will not be relevant, but as with science in general, there are threats to the geographical information sciences from politicians, propaganda, pseudo-science — and VGI.

- **Monopolies.**

Most official mapping agencies are inherently monopolistic, though they could get competition from other tiers of government and/or peer agencies, particularly in a country where there is poor political leadership, or the agency lacks the resources required or credibility. Such a monopoly can lead to high prices and products that don’t meet the needs of users. It can also lead to conceited arrogance and not recognising changes in technologies and markets and could render the agency marginalised, irrelevant or even redundant.

- **Sloppy work.**

This is unlikely to happen, unless staff are deployed for reasons of political allegiance rather than technical competence.

- **Unethical practices.**

These are only likely where there is political interference in the agency, such as manipulating statistics to cast the regime in a better light or censoring unpleasant data (eg: removing informal settlements from maps).

- **Plagiarism.**

For many national mapping agencies, this is considered to be a major problem, as individuals and companies transform the agency’s data and degrade their quality, while still labelling the data as being the agency’s data — and hence contravening the requirement for *truth in labelling*, see Section 6.2.

#### 4. User-generated content and volunteered geographical information

##### 4.8.5 Beyond traditional scientific media

A consequence of the bean counters reducing science to a numbers game of publications, citations and the like is that it becomes self-perpetuating, because “those who are highly rated by the system are frequently asked to rate each other and others; they are unlikely to want to change a system that gave them their status. Administrators often act as if only numbers count, a probability because their own evaluators do the same” [Parnas 2007]. Authors then need to trade off the opportunity costs of pursuing publication in journals preferred by bean counters, which often have high rejection rates and long lead times, with publishing quicker in other journals: earlier publication seems to yield more citations [Sekercioglu 2013].

Priem [2013] is naïvely optimistic about “the journal and article ...being superseded by algorithms that filter, rate and disseminate scholarship as it happens”, but he is promoting his product for doing this<sup>23</sup>. Nevertheless, he does make some interesting suggestions for the next decade, outlined below, as scientists are already using a diverse array of outputs.

- The **sheer volume** will overwhelm the traditional model of peer review.
- **Dissemination** will be decentralized, interoperable, open and diverse with evolving standards, and it will also include electronic conversations, data collection, premature results (*open-notebook science*) and shared analysis and description.
- **Review** will be “through the aggregated judgements of expert communities, supporting both rapid, fine-grained filtering and consistent, meaningful evaluation” — hopefully, this will make the review more rigorous. However, this could still fall into the trap of proof by repeated assertion and some of the problems described in Section 4.8.3 concerning traditional media, such as sloppiness, cliques and obsequiousness.
- **Alternative metrics** will provide many more options for **certification**, from which scientists can select; there are already companies providing services to support these new types of certification.
- **Filtering** is obviously widely used already (eg: Google Scholar), but this will expand to “personalized recommendation engines ... [to produce] a bespoke, curated stream” (but hopefully not a filter bubble).
- Finally, Priem [2013] feels that the bean counters will accept **new reward structures** for scholarship, such as the “impact of their diverse products” (a terror to which we are already subjected in South Africa<sup>24</sup>).

Indeed, Brembs *et al* [2013] conclude that not only is it poor scientific practice to use the impact factor of journals as an assessment criteria, but *any* form of journal ranking would have the same, negative impact. They would even abandon journals altogether as

<sup>23</sup>Essentially, an advertorial in *Nature*!

<sup>24</sup>Any measurement in the hands of bean counters and politicians without domain knowledge or experience is inevitably used terribly.

#### 4. User-generated content and volunteered geographical information

---

a means of communicating science, and rather use a “*library-based scholarly communication system*”, archiving and making accessible after peer review, papers with their associated software, raw data and metadata, exploiting modern information technology. Such a system would also include “scientifically-tested metrics accruing reputation in a constantly improving reputation system” [Brembs *et al* 2013]. Specifically, this would entail bringing “the discover functions of Amazon or eBay, the social networking functions of Facebook or Reddit and [of] course the sort and search functions of Google — all technologies virtually every scientist uses regularly for all activities but science” to provide evaluation options that are more scientific than journal impact factors (which actually are “negotiated, irreproducible and unsound”) [Brembs *et al* 2013].

Brembs *et al* [2013] point out that such a *library-based scholarly communication system* would still allow professional (and paid) editors to select worthy publications for inclusion in anthologies, digests and the like. The editors would obviously compete with one another, to persuade scientists and other paying customers that based on their successful track records of finding the most important discoveries or most interesting perspectives, their publications should be bought and serials subscribed to.

## 4.9 Citing

### 4.9.1 Citing user-generated content

Bibliographically, should the time the blog was posted, together with the day, month and year, be recorded in the citation? The problem is that some bloggers update their blogs occasionally, based on feedback from readers. The blogger might acknowledge the details of the update in a comment on their own blog — or they might not. The result is that when referencing such user-generated content it would probably be useful to identify the version being cited — particularly as the most interesting part of the blog might be the most controversial, and hence likely to be updated by the blogger as their opinion evolves in response to feedback. Similarly, a blog can be *remixed* (an alternative version of a text) or combined with other blogs in a *mashup* (content selection and aggregation from several sources). Such evolution of thought and the consequent updating of the written record is not unique to electronic media, of course, but is well understood, controlled and citable in the printed media — different editions of a book, or corrigenda, amendments, retractions or withdrawals of articles in journals, etc.

The BBC News Web site is a good example of maintaining appropriate metadata about its articles, providing a “signature” block at the end of the article containing the date, time and URL, such as the following for the news item by Fildes [2009]:

Story from BBC NEWS:

<http://news.bbc.co.uk/go/pr/fr/-/2/hi/technology/8209172.stm>

Published: 2009/08/19 12:23:26 GMT

© BBC MMIX



---

#### 4. User-generated content and volunteered geographical information

---

Even better, though, is the Web site of the British newspaper, *The Guardian*, which has for each article a pop-up window labelled *Article history*, providing metadata such as the following for the report by Cross [2007]:

**About this article**

**Close**

**Free our data: Address database plan finally abandoned**

This article was first published on [guardian.co.uk](http://guardian.co.uk) at 00.12 BST on Thursday 7 June 2007. It appeared in [the Guardian](#) on [Thursday 7 June 2007](#) on p3 of the [Technology news & features](#) section. It was last updated at 00.12 BST on Thursday 7 June 2007.

The underlined text provide hyperlinks: the first two to the main page for The Guardian's Web site, the third to the Main Section of the issue of Thursday, 7 June 2007, and the forth to the Technology Guardian section of the same issue.

As with the incremental updating and versioning of geospatial data (see Section 2.8), the bibliographic equivalent is needed for identifying, tracking and managing the incremental updating and versioning of user-generated content on the Internet (blogs, wikis, podcasts, etc) so that subsequent citations of that user-generated content remains valid. Part of the problem is identifying the original version of an article, because when searching for it on the Internet one gets presented with all the variations (including remixes and mashups of the article) by the search engine, but without sufficient metadata to identify the date of creation for all the alternatives or anything showing how they link together (eg: a graph or a dependency diagram). It is rather tedious to wade through each and every link proffered — one would have hoped that the search engine would eliminate such drudgery!

One example of this I experienced was trying to determine the original article where Lawrence Lessig first expressed concern about the Internet becoming “read only” (ie: controlled by corporations exploiting copyright law to sell content), as opposed to “read/write” (ie: where users create and recreate (or remix) content). It would appear that the original was an article published on FT.com on 28 December 2005 [Lessig 2005] (which might indicate that it first appeared in the newspaper, Financial Times, as an opinion piece), but when searching for it on Google at around 19:00 on 7 July 2009, the FT.com article did not appear in the first 500 items<sup>25</sup> offered by Google — I actually found the FT.com article through a review of it by Richard MacManus published on ReadWriteWeb on 17 January 2006 [MacManus 2006]. Hence, the search engine alone proved to be inadequate.

At the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference (incorporating the GISSA 2008 Conference), in Cape Town in October 2008 [Coetzee *et al* 2008c], Ed Parsons, the Geospatial Technologist at Google, claimed that with Google's search engine one did not need metadata. He then claimed to “prove” this by taking the audience through an example of using the metadata from the United States Geological Survey (USGS), a pioneer in recording metadata for geospatial data and in developing metadata standards. In this rather facile example, he attempted to use the USGS metadata to find the Grand Canyon, and “failed”. Obviously, he was deliberately misusing

---

<sup>25</sup>By when I got bored looking!

#### 4. *User-generated content and volunteered geographical information*

---

the metadata because when searching for a specific geographical feature by name one should not use the metadata but rather use a gazetteer, which is a geographical index, typically listing geographical names and their coordinates, such as is found in an atlas. An online example is the Official South African Geographical Names System [South African Geographical Names Council (SAGNC) 2016], which can be used to search for both old and new names and provides details such as the type of place, the date the name was approved, and even for some, a recording of the correct pronunciation of the name. However, a multilingual environment and traditional, colloquial and deprecated names will cause problems whether using metadata and a gazetteer, or searching.

One would use the metadata to find types of data sets (eg: landscape features) conforming to certain requirements (eg: currency). Within those data sets identified, one could search for specific features. The example given in Section 2.3.2 above of the paper allegedly from 1729 on spaceborne synthetic aperture radar [Armenakis *et al* 1729], illustrates the weakness of relying on only the search engine, which is dependent on whatever the relevant author(s) made available and how the search engine parsed the content.

##### 4.9.2 Citing data and repositories

While it is common for papers in the literature to cite the software used (typically, to the vendor's Web site or user manual), it is not yet common, unfortunately, for papers to cite the data sets or repositories used. The Committee on Data for Science and Technology (CODATA), of the International Council of Science (ICSU) established a joint Task Group on Data Citation Standards and Practices, with the International Council for Scientific and Technical Information (ICSTI). The Task Group aims at addressing the technical issues of interoperability and facilitation of re-use; the disparate needs of different scientific disciplines; the institutional and financial issues; sustainability; persistent identifiers for data sets; legal and intellectual property rights issues; and socio-cultural and community norms [CODATA Task Group on Data Citation Standards and Practices 2014]. The Task Group produced a report on the then state of practice, policy and technology for citing data [CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013]. They identified 10 "first principles" for data citation which they offered as guides:

1. **Status of Data:** Data citations should be accorded the same importance in the scholarly record as the citation of other objects.
2. **Attribution:** Citations should facilitate giving scholarly credit and legal attribution to all parties responsible for those data.
3. **Persistence:** Citations should be as durable as the cited objects.
4. **Access:** Citations should facilitate access both to the data themselves and to such associated metadata and documentation as are necessary for both humans and machines to make informed use of the referenced data.
5. **Discovery:** Citations should support the discovery of data and their documentation.

---

#### 4. *User-generated content and volunteered geographical information*

---

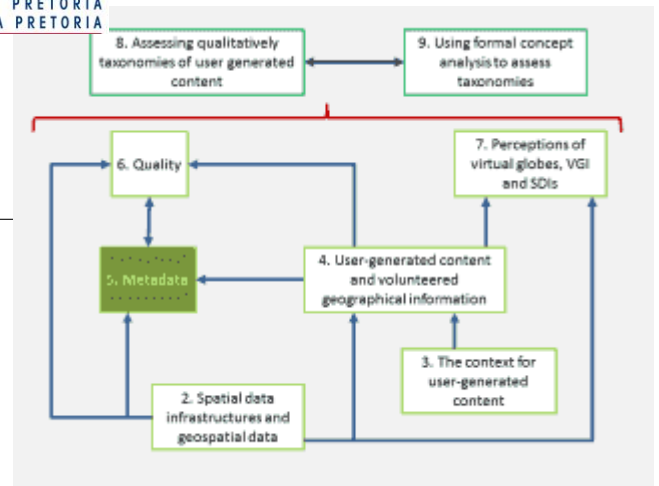
6. **Provenance:** Citations should facilitate the establishment of provenance of data.
7. **Granularity:** Citations should support the finest-grained description necessary to identify the data.
8. **Verifiability:** Citations should contain information sufficient to identify the data unambiguously.
9. **Metadata Standards:** Citations should employ widely accepted metadata standards.
10. **Flexibility:** Citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data across communities [CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013].

### 4.10 Summary and looking ahead

This chapter has provided details of user-generated content, citizen science, volunteered geographical information crowd sourcing, neogeography, the validity of using user-generated content in scholarly research, the quality of the traditional media, and citing user-generated content, data and data repositories.

Chapter 5 provides details of metadata. Specifically, it looks at the categories of metadata; the relationships between metadata and quality, and metadata and product specifications; and the definitions, aspects, encoding, tools, standards and limitations of and for metadata. It also compares metadata to searching and linked open data (LOD) and considers metadata for VGI.

\*\*\*\*



## Chapter 5

# Metadata

### 5.1 Overview of the chapter

Chapter 2 discussed spatial data infrastructures and geospatial data and Chapter 4 discussed user-generated content and volunteered geographical information. This chapter draws on them to present a detailed discussion of **metadata**, introduced in Section 2.7. Metadata is closely coupled with *data quality*, discussed in detail in Chapter 6, and both are key characteristics of VGI repositories, which are assessed in Chapters 8 and 9. Further, the most common objections raised against VGI are perhaps the uncertainty over the quality of the VGI and of the documentation of the data, that is, the metadata (eg: Cooper *et al* [2010a]). Specifically, this chapter covers the following.

- Section 5.2 provides an introduction to metadata and some *definitions* of metadata.
- Section 5.3 takes this further, describing aspects of metadata such as its relationship to *quality*, and *active* and *passive metadata*.
- Section 5.4 describes the benefits of *encoding metadata*.
- The relationship between a product *specification* and metadata is described in Section 5.5.
- Section 5.6 describes *tools* for capturing metadata.
- The different *categories* or types of metadata are described in Section 5.7.

- Section 5.8 provides some principles for *standards* for metadata and then describes various such standards.
- Section 5.9 discusses some *limitations of metadata*.
- There are those who consider *searching* to be superior to metadata, which is discussed in Section 5.10.
- Section 5.11 discusses metadata and *linked open data (LOD)*.
- Finally, Section 5.12 discusses VGI and metadata.

The key contribution that I have made that is presented in this chapter is:

- A comprehensive summary of the different *categories* or types of metadata, which are described in Section 5.7.

Hopefully, this work contributes towards a new and improved understanding of VGI, its metadata and also its usability. Other authors have also contributed in this regard, for example, Elwood *et al* [2012] who aimed “to frame the crucial dimensions of VGI for geography and geographers, with an eye toward identifying its potential in our field, as well as the most pressing research needed to realize this potential”.

This Chapter draws on my involvement in various metadata initiatives over the last 30 years, such as the South African National Exchange Standard (NES)<sup>1</sup> [Clarke *et al* 1987, 1988; Cooper 1988*a*, 1986, 1987*b,a*, 1989*a,b*; Cooper & Clarke 1991; Cooper 1993, 1997]; the research and publications of the International Cartographic Association’s Commission on Geoinformation Infrastructures and Standards, and its predecessors [Moellering 1991; Cooper & Clarke 1991; Moellering & Hogan 1997; Cooper 1997; Cooper & Nielsen 2000; Moellering *et al* 2005; Cooper & Gavin 2005]; the metadata standards of ISO/TC 211, *Geographic information/Geomatics* [ISO 19115 2003; ISO 19115-1 2014; Cooper 2007, 2009*a*]; the South African profile of ISO 19115 [SANS 1878 2005]; South Africa’s Committee for Spatial Information (CSI) [Harvey *et al* 2012; Cooper 2013]; and various projects at the CSIR, including for the then Department of Water Affairs [Olivier *et al* 1990]; for Statistics South Africa (StatsSA)<sup>2</sup> [Lukhwareni *et al* 2005*a*; Cooper 2005]; for guidelines for data content standards for Africa<sup>3</sup> [Cooper *et al* 2005]; and for the Chief Directorate: National Geospatial Information, chapters on standards for fundamental geo-spatial datasets [Coetzee *et al* 2014] for a book of guidelines for the Mapping Africa for Africa initiative [Clarke 2014].

Some of my PhD research was supported by a collaborative project with the University of Pretoria and the Wrocław University of Environmental and Life Sciences<sup>4</sup> [Cooper 2011*c*; Cooper *et al* 2011*a*].

---

<sup>1</sup>While not described explicitly as metadata in the technical specification, many of the constructs in NES were for metadata.

<sup>2</sup>Two service level agreements funded by the CSIR and StatsSA, in terms of the Memorandum of Understanding covering the CSIR’s participation in the National Statistics System (NSS).

<sup>3</sup>Initiated by the EROS Data Center of the United States Geological Survey (USGS/EROS) and EIS-Africa, with funding from the United States Agency for International Development (US AID) and the CSIR.

<sup>4</sup>Funded partially by a Joint Research Grant under the SA/Poland Agreement on Cooperation in Science and Technology.

## 5. Metadata

---

### 5.2 Definition of metadata

As discussed above in Section 2.3.4, digital geospatial data are complex, and abstract concepts describing the geospatial data have to be made concrete in a GIS to be rendered to data structures and code. All digital data attempt to model and describe the world. Such models are meant to reflect the real world (as it is), the imaginary world (as planned, as it might have been, as forecast, etc), or some combination of the real and imaginary worlds. They are always only an abstraction of reality; they are always partial (reflecting the conscious and unconscious biases of the compilers); and they are not complete. Any geospatial data set is always only just one of many possible ‘views’ of the world — they cannot be an exact duplication of the world. For any data set, some things are approximated, some things are simplified and some things are ignored. Hence, there can never be perfect, complete and correct data! So, to ensure that data are not misused, all the assumptions made in creating a data set and all of the limitations should be documented fully — as *metadata*<sup>5</sup>.

For any system to be viable, its components, products, services and their underlying data must be well documented. The metadata enables products, services and data to be discoverable and comparable. Metadata is also a necessary precondition for analysis, data conflation, quality assessment and methodological evaluation. Many users complain of struggling to find data sets (see Section 7.5 and Goodchild [2007b]; Craglia *et al* [2008]), but the problem is not the availability of data. There are vast amounts of free or very cheap data that are available. The real problems are finding the data, assessing the suitability of the data and integrating data together: for all of these metadata is critical.

Traditionally, metadata has been defined as being “data about data” (eg: in ISO 19115 [2003]), but metadata is more than just that. Metadata also describes processes, services, systems, etc. ISO makes the definitions in their standards available through the Concepts Database [ISO 2016b]. ISO’s Directives provide stringent rules regarding definitions [ISO/IEC 2011b], though they are not always adhered to! Some definitions of metadata from various standards are given below verbatim, including capitalization and punctuation.

- “*information about a resource*”, where a resource is an “*identifiable asset or means that fulfils a requirement*”, such as a data set, data set series, service, document, initiative, software, person or organisation [ISO 19115-1 2014].
- “*data that defines and describes other data*” [ISO/IEC 11179-1 2004].
- “*Additional data associated with the image data beyond the image data*” [ISO/IEC 15444-2 2004].
- “*Data about data or data elements, possibly including their data descriptions, and data about data ownership, access paths, access rights and data volatility*” [ISO/IEC 2382-17 1999]<sup>6</sup>.

---

<sup>5</sup>The introduction to ISO 19115 [2003] contains similar wording, because I contributed some of that text.

<sup>6</sup>Please note that this is not a metadata standard, but it provides database terms and definitions in English and French, for use by other standards.

- “The information and documentation which makes multimedia data understandable and shareable to users over time” [ISO/IEC 15938-5 2003].
- “data describing context, content and structure of records and their management through time” [ISO 15489-1 2001].
- “structured data about data, including data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation” [ISO 11620 2008]<sup>7</sup>.

Surprisingly, the well known metadata standard, ISO 15836:2009, *Information and documentation — The Dublin Core metadata element set*, does not include a definition of *metadata*! Technically, this means that as used in the standard, the term is self-explanatory, commonly used and not interpreted differently in different contexts [ISO/IEC 2011b]. As the standard is meant to have a broad scope (providing metadata for a *resource*, which is “anything that might be identified” [ISO 15836 2009]), the writers of the standard might have felt that providing a definition for *metadata* would have limited its scope.

As can be seen, these definitions given above have much in common, even though they generally reflect their respective domains. Combining them, metadata tends to be *structured* and relates to the *resource* (data set, service, etc) it describes, but goes beyond to make the resource *understandable* and *shareable* into the future (ensuring *longevity* is a critical aspect of metadata — particularly for the producer of the resource for their own purposes) and to *describe aspects* such as the content, context, functionality, data structures, data descriptions (eg: classification), ownership, access mechanisms, legal issues (eg: access rights) and data volatility (ie: currency and maintenance cycle). Metadata *facilitates* the administration, management, maintenance, discovery, access, retrieval, assessment, use and preservation of the resource. Note that a resource described by metadata, does not have to be a digital object or Web service, but could also be a physical object, such as a paper map, a log book, a specimen<sup>8</sup> or an artefact.

The following definition is from a text book on spatial databases:

*“A formally structured and documented collection of information about data that reveal minimally what is in the data, where the data originated from, who produced them, when they were produced and modified, why they were produced, and how the data can be obtained. ‘Data’ in this definition refer generally to a database, a data set, and even a data element. Hence, metadata include basically everything about the data except the data themselves” [Yeung & Hall 2007].*

Clearly, this definition is similar to the combination of the definitions taken from various ISO standards, as described above. Although not explicit in any of these definitions, a key aspect of metadata is to try to *ensure the correct use* of the resource, or to try to prevent the misuse of the resource. Part of this would be covered by the principle of *fitness for use* of data quality, as described in Section 6.2, but metadata describes more than just quality.

<sup>7</sup>Please note that this standard specifies library performance indicators and while it defines *metadata*, it only mentions it in passing under one of the indicators.

<sup>8</sup>Such as a fossil, eg: as recorded in the geospatial database of fossils of the Beaufort Group (Karoo Super-group) in various museum collections [Van der Walt *et al* 2010, 2015].

## 5. Metadata

An example of this would be using a data set in contravention of the copyright on the data, the specifications for which would be included in the metadata for a geospatial data set using *MD\_LegalConstraints.useLimitation*, as described in ISO 19115 [2003].

### 5.3 Aspects of metadata

Metadata as a premise covers many different concepts, including *quality* (see Chapter 6), that is, metadata can document quality. Metadata and quality should not be limited to describing only a resource or its elements, which would in any case include the processes that created the resource — its *lineage*. They should also describe the context in which the resource is created, such as the responsible organisation itself; access to the data set; the series for which the data set forms a part; etc. Clearly, there is a danger in requiring too much metadata because of the costs of acquiring the metadata. However, I know from my roles in ISO/TC 211 that the project teams for ISO 19115 [2003] and related standards, for example, include experts who generate vast quantities of metadata daily (such as from satellite imagery).

There is also some confusion over terminology, and this chapter and the next are an attempt to reconcile the different models of *metadata* and *quality*. *Quality* is not a subset of *metadata*, because not all of the *inherent quality* of a resource gets documented, see Section 6.3, but the documentation of quality is part of metadata. Similarly, *metadata* includes aspects unrelated to *quality*, such as data formats. The quality of metadata is discussed in Section 5.12.

As is detailed in ISO 19115 [2003] and ISO 19115-1 [2014], metadata can have coarse or fine *granularity*, that is, describe the likes of a dataset series (eg: South Africa's 1:50 000 national mapping series), a platform (eg: South Africa's satellite, Sumbandilasat), a sensor (eg: a weir gauge), a Web service, a dataset (eg: a 1:50 000 map sheet), an individual record, or an individual field. Metadata allows a producer to describe a resource fully so that users can understand these assumptions and limitations and evaluate the resource's applicability for their intended use.

Metadata can be *passive* or *active*.

- **Passive metadata.**

This focuses on documentation for use by humans, such as for data discovery and assessment, and free text is quite acceptable. This approach has been adopted by the *Dublin Core* metadata standard [ISO 15836 2009], for example (see Section 5.8.8). One criticism of this approach is that while it makes it easier to produce compliant metadata (because of the variety allowed by free-text fields), it makes it difficult to process the metadata — resulting in some cynics describing such standards as “write only” or “input only”!

- **Active metadata.**

This focuses on driving systems and processes, such as the automated selection of candidate data sets or triggering alerts about system failures, and hence needs to be encoded (see Section 5.4), because of the complexities of interpreting free text

automatically. ISO 19115-1 [2014] uses encoding extensively, for example (see Section 5.8.5).

Metadata harvesting is aided by active metadata, but a harvesting service should be able to deal effectively with passive metadata, to provide a better discovery service.

From the perspective of the user, metadata and quality can also be *explicit* (as encoded in the metadata accompanying the resource) or *inferred*, based on the user's experience, the reputation of the resource supplier, through not reading or understanding properly the metadata actually supplied, by ignoring the metadata, or wishful thinking. This is different from any inference that might be done by a producer when creating the metadata, as discussed in the item on *implicit metadata*, see Section 5.7.2.

## 5.4 Encoding metadata

Metadata can be recorded in a variety of forms, from free text to being rigorously encoded. The metadata could be just a word-processed document (as happens in many organisations), could be embedded within the spatial data set the metadata describes, or could be stored in a special database or registry for metadata. Metadata can include explanatory notes, manuals, research papers and other resources that describe the data, products, services, processes or systems — and even experiences with using them. The metadata content could be arbitrary, or could conform to a standard. Key issues are:

- Being able to ensure that the metadata remains with what it describes (whatever the granularity), even after being distributed widely;
- Ensuring that the metadata is available to be located by the relevant search and harvesting services;
- Ensuring that the metadata can be processed automatically, as appropriate;
- Supporting multiple languages, character sets and cultural contexts;
- Aggregating or integrating metadata when data sets are combined;
- Ensuring that the metadata does not place too high a demand on the user's knowledge of the metadata elements, structure, etc; and
- Ensuring that the tools and standards for metadata encourage rather than discourage the capturing of metadata.

There is a variety of metadata standards available, with their principle differences being the application areas they cover, their breadth, their depth (ie: the amount of detail for which they cater), and their encoding. Most metadata standards include limited encoding of their elements, preferring to cater for free text. As discussed in Section 5.3, while this makes it easier for the producer capturing the metadata, it makes it difficult for the user comparing metadata, and very difficult to implement any automated processes based on the metadata. Several different metadata standards are discussed below in Section 5.8. Table 5.1 illustrates the difference between encoded metadata (selected

## 5. Metadata

from ISO 19115-1 [2014]) and free-text metadata (selected from Dublin Core [ISO 15836 2009]). Encoding metadata (particularly through code lists or enumeration) makes it easier to implement multilingual interfaces to the metadata, to translate metadata between languages, and to export the metadata in formats conformant to other metadata standards. Encoding the metadata also facilitates active metadata.

Table 5.1: Encoded and free-text metadata

Type	Encoded [ISO 19115 2003]	Free-text [ISO 15836 2009]
contributor	CI_Citation.title = CSIR. CI_Citation.citedResponsibleParty.role = <i>contributor</i> . CI_Citation.presentationForm = <i>documentDigital</i> .	contributor = CSIR. <i>It is a digital document.</i>
Coverage	EX_Extent.EX_GeographicExtent.extentTypeCode = 1. EX_Extent.EX_GeographicExtent. EX_GeographicBoundingBox. westBoundLongitude = 10.0. EX_Extent.EX_GeographicExtent. EX_GeographicBoundingBox. eastBoundLongitude = 20.0. EX_Extent.EX_GeographicExtent. EX_GeographicBoundingBox. southBoundLatitude = -40.0. EX_Extent.EX_GeographicExtent. EX_GeographicBoundingBox. northBoundLatitude = -30.0.	coverage = 10 degrees E, 20 degrees E, 30 degrees S, 40 degrees S.
Description in mixed languages	MD_Identification.abstract = PT_Locale.language = <i>eng</i> , PT_Locale.country = ZA, PT_Locale.characterEncoding = 3, <i>This draft map shows the locations of the campuses of the Technical Vocational Education and Training Colleges that we have been able to identify from the Web pages of the Colleges, virtual globes and other Web sites. There probably are errors in the data..</i> PT_Locale.language = <i>afr</i> , PT_Locale.country = ZA, PT_Locale.characterEncoding = 4, <i>Hierdie kaart bewys die liggings van die Tegnies Beroepsonderwys en Opleiding Kollege terreine wat ons kon uitken van die Web-bladsye van die Kolleges, virtueel aardbolle en ander Web tuiste. Daar is seker foute in die data..</i>	description = <i>This draft map shows the locations of the campuses of the Technical Vocational Education and Training Colleges that we have been able to identify from the Web pages of the Colleges, virtual globes and other Web sites. There probably are errors in the data. Hierdie kaart bewys die liggings van die Tegnies Beroepsonderwys en Opleiding Kollege terreine wat ons kon uitken van die Web-bladsye van die Kolleges, virtueel aardbolle en ander Web tuiste. Daar is seker foute in die data.</i>

Referring to Table 5.1, each field would be identified in practice in the usual way (eg: tags in XML) and only the text shown in italics in the table would actually be included in the field. For the ISO 19115 [2003] examples, some of the intermediate metadata elements have been omitted, to improve clarity. An *EX\_Extent.EX\_GeographicExtent.extentTypeCode* with a value of 1 means the metadata is for the area indicated (0 means the metadata

is for everywhere but the area indicated). Normally, default values would be set for *PT\_Locale.language*<sup>9</sup>, *PT\_Locale.country*<sup>10</sup> and *PT\_Locale.characterEncoding*<sup>11</sup>, and hence could be omitted for the default language in the example above. On the other hand, some might prefer to retain them for all free text fields in the metadata, to reduce any implied cultural or linguistic bias. For *PT\_Locale.characterEncoding*, 3 is for 7-bit US-ASCII [ANSI 1986], which does not have the accents needed for Afrikaans, and 4 is for ISO 8859-1 [1998], which is the 8-bit encoding for Western Europe that includes the accents used in Afrikaans. The *PT\_Locale.language* and *PT\_Locale.country* combination allows one to distinguish between the Portuguese of Brazil and Portugal, or the German of Austria, Germany and Switzerland, for example.

By encoding the information in business reports, as opposed to just having blocks of text, the *eXtensible Business Reporting Language* (XBRL) allows for the easy translation of such reports into other languages and the automated processing of business and financial information: reducing errors, permitting automated checking and speeding up the handling of the information. More importantly, though, it supports the various accounting standards, banking regulations and reporting regimes around the world, and facilitates performance benchmarking and analysis by consumers of the information (investors, analysts, regulators, etc). XBRL also reduces the scope for ambiguous or manipulated reporting by companies, as all the fields are well defined [XBRL International 2016]. These benefits can obviously apply to encoded metadata, as well.

## 5.5 Metadata and specifications

Metadata as a concept is increasingly being used to describe more than just data — that is, the definition is being broadened to be “data about anything”. This has value, in that the tools, techniques and methodologies developed for metadata can be applied more widely. In any case, metadata has always been applied more widely: when documenting the *lineage* of a data set, one describes processes, and when documenting the *distribution information* for a data set, one documents organisations, pricing policies and the like. However, care must be taken that the concept of metadata is not ‘cheaper’ by overly expanding its scope, resulting in it losing its value for describing a resource usefully.

To evaluate the *fitness for use* of a data set (or product or service, etc), a user needs to compare the data set’s metadata against their *data product* or *user requirement specification* (or any other type of specification), that is, against the systematic description of the data that the user needs. Unfortunately, it is possible to confuse a *specification* with *metadata*. The specification could be matched against the metadata for one or more candidate solutions to determine the extent to which each solution satisfies the specification — that is, their *fitness for use*. It is probably better to separate the specifications from the metadata, so that it is explicit as to which is being dealt with.

<sup>9</sup>Taken from ISO 639-2:1998, *Codes for the representation of names of languages — Part 2: Alpha-3 code*.

<sup>10</sup>Taken from ISO 3166-1:2013, *Codes for the representation of names of countries and their subdivisions — Part 1: Country codes*.

<sup>11</sup>Taken from the IANA Character Set register [IANA 2016].

## 5. Metadata

---

Generally, there would be only one candidate solution when that solution was developed by (or on behalf of) the user in response to the specification. There could be several candidates, or even many, when the solution is being sought from existing ones in the market. A rigorous adherence to a good metadata standard facilitates such assessment, which is why data producers and users need to invest much time and effort into understanding metadata and quality. The product specification could be the *raison d'être* for the data set (ie: the data set was created to meet the needs of the product specification), or it could be developed afterwards for data discovery (ie: searching for candidate data sets for one's application).

Using the same basic elements in both the metadata and the specification facilitates matching them against one another, particularly if the elements are encoded, as opposed to just being free text. It would also enable automating some or all of the comparisons, which would be particularly useful for data discovery, when assessing a large number of candidates. For example, if the specification requires that the data have a security classification of 'unclassified' and are available on a CD-ROM, having both "security classification" and "medium type" elements in both the specification and metadata, and with the same code lists, would make it easy to compare the candidate metadata against the specification. The metadata standard ISO 19115 [2003] and ISO 19131:2007, *Geographic information — Data product specifications*, have been so harmonized. ISO 19131 defines a data product specification as a "detailed description of a data set or data set series together with additional information that will enable it to be created, supplied to and used by another party". To describe a product specification, ISO 19131 uses the metadata elements defined in ISO 19115, which obviously facilitates the direct comparison of a data product specification to the metadata of candidate data sets. ISO 19131 includes the application schema, spatial and temporal referencing systems, quality and data capture and maintenance processes. Simplistically, the matching process could be expressed as the following equation, where  $M$  is the proportional match of the supplied data set to the data product specification and  $1 = \text{Perfect match}$  and  $0 = \text{Complete mismatch}$ .

$$\text{ISO 19131} - \text{ISO 19115} = 1 - M \quad (5.1)$$

However, not every metadata element will necessarily have a corresponding specification element, and vice versa. Further, not every metadata element is relevant for an assessment of fitness for use, as many indicate how to interpret the data, how to transform the data set to meet the user's requirements, or other things — examples include the character set used to encode the metadata, contact details for the maintenance authority, and the metadata standard used.

Currently, while there has been much focus on metadata and metadata standards, there has not been much focus on product specifications or standards for them. However, the Committee for Spatial Information (CSI) is developing a Data Collection Project Register (DCPR) for data custodians to notify their intention to capture a geospatial data set [Siebritz & Fourie 2015], and DCPR will probably be based on ISO 19131 [2007].

Unfortunately, many users do not develop good product specifications before commencing their search for a suitable data set, with the result that every data producer probably

needs to field on a daily basis, vague queries for data! This is undoubtedly because of a lack of knowledge on the part of the users, rather than laziness or incompetence. Trawling through the metadata of candidate data sets will then help the user determine what their product specification really should be.

## 5.6 Metadata tools

Unfortunately, implementing metadata is often not treated seriously, apparently because it is perceived to be tedious and expensive to capture and because there is no perceived value to the capturer: “*Metadata collection is dull, expensive, and time-consuming*” [Sundgren 1995]. Further, errors can be made in capturing the metadata, such as typographical errors (eg: omitting or transposing letters), scanning errors (eg: when scanning a paper map), data conversion errors (ie: an error in the software reading or writing the metadata), find-and-replace errors (eg: invalid error correction) and metadata errors (eg: not updated) [Beall 2006b]. In collaborative projects, it can be useful having deadlines to get metadata documented and submitted [Ellul *et al* 2012]. Further, the difficulty is not just with capturing metadata, but also with presenting it meaningfully for the user and the usability of the metadata [Poore & Wolf 2013].

These are legitimate concerns over metadata, which makes it crucial to have suitable tools that automate metadata capture as much as possible. For example, referring to ISO 19115 [2003, p 16], the following core metadata could be captured automatically, as discussed:

- *Dataset responsible party* (`MD_Metadata > MD_DataIdentification.pointOfContact > CI_ResponsibleParty`): all the relevant details of the organisation should be stored and maintained on the system as a matter of course, so it would be easy to use such data to populate these fields.
- *Geographic location of the dataset* (`MD_Metadata > MD_DataIdentification.extent > EX_Extent > EX_GeographicExtent > EX_GeographicBoundingBox`): the extent of a data set is normally calculated by a GIS and stored in the data set, to facilitate using the data set. Otherwise, it is trivial to calculate automatically.
- *Dataset language* (`MD_Metadata > MD_DataIdentification.language`) and *Dataset character set* (`MD_Metadata > MD_DataIdentification.characterSet`): these will generally be default values but if they are not, they will need to be recorded in any case, so that the data set can be used.
- *Reference system* (`MD_Metadata > MD_ReferenceSystem`): this will need to be recorded in the system, so that the data set can be used, particularly if it is being used with other data sets covering the same geographical area.
- *Metadata standard name* (`MD_Metadata.metadataStandardName`) and *Metadata standard version* (`MD_Metadata.metadataStandardVersion`): these should be recorded in the system, because they determine what metadata can be made available.

Most decent GISs include tools for capturing metadata compliant to ISO 19115 [2003] and its related standards, and to the old Content Standard for Digital Geospatial Metadata

## 5. Metadata

---

(CSDGM) [FGDC 1998]. There are also stand-alone tools for capturing metadata, providing cross-walks between metadata standards [FGDC 2006] and converting metadata from one format to another. The American Federal Geographic Data Committee (FGDC) maintains a Web page with some details of metadata tools [FGDC metadata 2016].

### 5.7 Categories of metadata

#### 5.7.1 Overview

There are different types or *categories* of metadata and for these, metadata can be created manually, semi-automatically (eg: through a highly structured user interface) and automatically (eg: through defaults, or reading system parameters such as date, time and the details of the user logged in). Metadata can also be inherited, derived and/or integrated from the various input data sets and processes. Metadata can also be inferred from the resource or other metadata, such as inferring the location where a photograph was taken from the social context, annotations and other metadata [Davis *et al* 2004].

One type of organisation that has invested in understanding metadata and capturing metadata for their products is the national statistical agency. The types of metadata that they use are discussed in Section 5.7.2. In Section 5.7.3, I present several types of metadata not used by the statistics agencies, drawn mainly from my experiences within ISO/TC 211. Section 5.7.4 discusses the grouping of metadata categories.

#### 5.7.2 Categories used by statistical agencies

There is a variety of international groups dealing with the enterprise architecture and related aspects of statistical agencies, that are led primarily by the United Nations Economic Commission for Europe (UNECE), Eurostat (the European Union's statistical office) and the Organisation for Economic Co-operation and Development (OECD). Other UN agencies are also involved, such as the United Nations Industrial Development Organization (UNIDO).

One of these UNECE/Eurostat/OECD groups is the Work Session on Statistical Metadata (METIS), that has been preparing a Common Metadata Framework (CMF), the Generic Statistical Business Process Model (GSBPM) and the Generic Statistical Information Model (GSIM) [United Nations Economic Commission for Europe 2016]. CMF is an active repository aimed at helping statistical agencies choose standards, models and approaches for their metadata systems. GSIM aims at providing common terminology and definitions for developing metadata systems and information management frameworks [METIS 2011].

The following categories of metadata have been gleaned from them [Lukhwareni *et al* 2005b; Malimabe & Jenneker 2010; United Nations Economic Commission for Europe 2016, 2009; Upadhyaya & Todorov 2009], but they do overlap as they adopt different perspectives of metadata: type, purpose, structure, manuals, etc. While aimed primarily at statistical products, they do also apply to data and services.

- **Definitional, descriptive, statistical or conceptual metadata.**

These are the identifiers and descriptors for statistical units, topics, classifications, coding schemes, concepts and definitions, vocabularies, ontologies, question modules, data items and quality measures. While this category includes descriptive text, the concepts are grouped into logical topics — which is effectively a classification, drawn from different study domains. This metadata can be context free (eg: the variable ‘income’ as a concept) or context sensitive (eg: ‘income’ collected in a particular survey).

This is the traditional category of metadata, tends to form the bulk of the metadata, and is considered to be the most important by some, particularly for enabling re-use. If a metadata standard has not been used, *definitional metadata* could also include definitions of the metadata elements, etc.

*Conceptual metadata* could also be considered separately, as a framework for describing the basic idea (concept) behind the metadata objects.

- **Procedural or methodological metadata.**

These describe the procedures by which data are collected, processed and reported, such as sampling procedures, sample sizes, collection methods, collection dates, deadlines, maintenance, editing processes and constraints (eg: embargo periods, particularly for sensitive national statistics, such as inflation).

- **Operational or quality metadata.**

These describe the results of implementing the procedures and methodologies, eg: measures of response burden, response rates, edit failure rates, status, weightings, versions, costs and other quality and performance indicators. In other words, the operational metadata is used to explain how the data was created or transformed. Conceptually, *operational metadata* could follow the *lineage* chain of the *sources* of each data item and the *methods* applied for its generation (see Section 6.5).

The *procedural metadata* describes the ideal process, while the *operational metadata* describes what happened in practice. The former can determine the *anticipated* or *planned* degree of confidence in results (effectively, giving an upper bound), while the latter can describe the *actual* degree of confidence in the results.

- **Systems, technical, control, distribution or administrative metadata.**

These describe the physical, unique characteristics of the data (ie: the *syntax*) for internal operations, concerned with the processes affecting data or products, rather than with the content. This metadata should be *active metadata*, to drive or facilitate automated workflow and operations, manage content and/or trigger alerts. Such metadata includes file layout, access paths, business processes, publication or dataset identifiers, dates (creation, update, release, archive, access, etc), file size, record lengths, formats, mappings between logical and physical names of files, dataset input flows, methods to access databases, table and column definitions, schemas and mappings of data. The *Statistical data and metadata exchange (SDMX)* standard, for example, is essentially for systems metadata, to enable the automated harvesting of national statistics from the Web sites of national statistics agencies [SDMX 2009].

## 5. Metadata

---

- **Discovery metadata.**

This is used by customers to identify, find and navigate through the information they need, such as releases, publications and data sets. This metadata includes the likes of title, scope, keywords, author, publication date, variable names and geographical areas. Such metadata provides a way of finding relevant data and content using search engines and indexes. *Discovery metadata* cuts across other types of metadata, particularly *definitional metadata*, and hence it is often synthesised from the other types of metadata, rather than captured explicitly as discovery metadata.

- **Dataset or survey metadata.**

This is the minimal metadata used to describe a dataset and its data structure, including access thereto, and update thereof, and broader aspects of the survey, such as information about the population described by the data. This appears to cut across other types of metadata (as with *discovery metadata*), providing some core set of metadata for a specific dataset or survey. The content might well be similar to that of *discovery metadata*.

- **Implicit metadata.**

This is used by UNIDO to describe metadata about one thing that is ‘transferred’ to another through other metadata, primarily *definitional metadata*. For example, the data for several classes could be combined but reported for only one of the classes as a surrogate for the combination, or a code from one domain for an object could be correlated with that from another domain [Upadhyaya & Todorov 2009]. As should be obvious, this first type of *implicit metadata* is to compensate for limitations in a classification scheme (or whatever other limitations there might be). The second is the equivalent of a database *join* between two classification schemes, or even just a classification scheme with two codes for each class, such as ISO 639-5 [2008] and ISO 639-6 [2009], which together provide three- and four-letter identifiers for the names of natural languages.

In practice, *implicit metadata* would go beyond these two types, to include the likes of search engine databases, *linked data* (registries, repositories, ontologies, aggregators, services, etc — the *Semantic Web*, see Sections 3.5 and 5.11), cross-referencing between different metadata files, and unravelling a chain of *lineage* records (to access the source data sets and then their metadata).

However, I should point out that there are those who are sceptical about the whole of the UNECE/Eurostat/OECD process regarding metadata: for example, see the anonymous and often vitriolic blog, “Information Systems in Official Statistics: What everyone should know about statistical metadata” [Anonymous 2011]. There do seem to be a lot of committees and groups from UNECE, Eurostat, OECD and related organisations that are dealing with metadata and some of them have been active for a very long time.

### 5.7.3 Other categories of metadata

The following categories of metadata appear to have been omitted by these agencies, though they might be implicit in some of the other categories:

- **Spatio-temporal metadata.**

This is probably included in *definitional metadata*, and would cover spatial extent, resolution, spatial representation (eg: vector or gridded), spatial rectification, spatio-temporal primitives (eg: a point and a date), spatial and temporal reference systems, topology and the like [ISO 19115-1 2014].

- **Organisation metadata.**

This would describe the organisations and people responsible for the whole life cycle of the data, products and/or services, including initiation, collection, publishing, maintenance and, of course, the metadata. A key problem with organisational metadata is providing persistent contact information and identifiers, because organisations change and people move around.

- **Policy metadata.**

This would describe the policies behind the whole life cycle of the data, products and/or services, including initiation, collection, publishing and maintenance. The policies will include legislation and business agreements, and cover issues such as maintenance, update cycles, versioning, application schemas, access, dissemination, portrayal, archiving, backups, standards, copyright and other rights, pricing, security, conformance testing and liability. There should also be a policy on metadata itself!

- **Use Metadata.**

This would document and manage user access, user tracking and multi-versioning information [Higgins nd]. There is enormous value for the producer in such metadata, for understanding the popularity of their offerings and for tailoring them for their markets. A complex aspect is that of *incremental updating and versioning* [Peled & Cooper 2004], see Section 2.8, to ensure that the producer keeps track of which user got which version of their data, so that they can advise each user explicitly on how changes to the base data will affect the georeferencing and other aspects of the user's value-added data. This problem becomes even more complex when data are distributed continuously (eg: satellite imagery — see the discussion on Figure 6.5) and/or piece-meal (eg: by cadastral parcel, when each is updated, created or deleted independently), rather than only occasionally in batches.

- **Preservation Metadata.**

This would document the actions undertaken to preserve a resource (digital or analogue), such as migrations and checksum calculations [Higgins nd] and the file plan [Schmitz & Cooper 2009]. This is essential for data curation.

- **Constraint metadata.**

This would document the legal, security, moral and other constraints there might be on accessing and using resources and the metadata about resources. Such constraints include copyright, patents, trademarks, licences, security classification and privacy. Constraints can be imposed by statute, regulations, contracts, agreements and even accepted practice. Constraints can also be time-dependent, such as an embargo date and time for releasing market-sensitive information, eg: inflation or GDP.

## 5. Metadata

- **Portrayal metadata.**

This would document how the resource could be rendered or displayed for human use, such as maps, graphs or tables. It would also include thumbnail graphics and other illustrations of or for a resource, including legends, organisation logos and constraint logos (eg: for Creative Commons [2016]).

- **Localisation metadata.**

This would cover the cultural and linguistic adaptability (CLA) of the resource and the metadata, such as the script, character set, language and language variant used. A good metadata standard (eg: ISO 19115-1 [2014]) would allow a mix of these, particularly for countries which have multiple official languages. A *locale* is a combination of language, character encoding, script, regional variations and other characteristics required to localise text. The intention also is to reduce the cultural and linguistic bias, such as the domination of the United States of America and American English in information technology.

- **Meta-metadata.**

This would be the documentation of the metadata itself, such as the metadata standard used and which version (including profiles and extensions); the scope of the metadata; how the metadata is distributed (embedded in the data set described, included in an associated file and/or available from a metadata registry); how the metadata is maintained; and the source of the metadata, such as the creator of the data set or a third party (using available sources, domain knowledge and expertise) [ISO 19115-1 2014; Duval *et al* 2002].

- **Metadata extensions.**

Nominally, a good metadata standard should not need extension! However, in practice it is necessary to cater for extensions. Firstly, any metadata standard needs to be constrained to ensure that it will actually be implemented. Secondly, it would be arrogant of the developers to assume that their standard has covered everything. Thirdly, extensions allow a standard to be used to a finer level of granularity within organisations and sub-units of an organisations, across disciplines and in area not previously imagined. On the other hand, catering for extensions can render the intended implementations of the standard as *write only*, that is, reduce interoperability.

In terms of the packages in ISO 19115-1 [2014], the *application schema topic category* would be covered by the *conceptual metadata* and *maintenance* would be covered by the *procedural metadata*.

### 5.7.4 Grouping categories of metadata

Potentially, all these metadata categories could be grouped, say into *descriptive*, *structural* and *administrative* metadata HLWIKI Canada [2015a], as follows:

- **Descriptive metadata:** documenting the actual *content* for identification, searching, retrieval, etc, eg: an abstract;

- **Structural metadata:** describing the layout, architecture and relationships between different parts of the resource, such as for navigating through the resource, eg: table of contents or page numbers; and
- **Administrative metadata:** describing technical aspects for managing and processing the resource, eg: confidentiality issues.

Duval *et al* [2002] recommend separating the *semantics* of the metadata from the *syntax* of the metadata, because of the diversity of data formats and the speed with which they change — while the semantics remain unchanged. Duval *et al* [2002] also point out that metadata can be *objective* (eg: custodian contact details, creation date and coordinate reference system) or *subjective*, eg: abstract, usage constraints and value judgements about the quality of the data set. What is also subjective is the selection of the metadata elements used and the amount of detail provided for each metadata element.

## 5.8 Standards for metadata

### 5.8.1 Principles for standards for metadata

Metadata standards aim at allowing producers to describe data sets and services fully and coherently, and at facilitating discovery, retrieval and use of data and services. A *metadata standard* can also help ensure uniformity in data collection, and improve data management, use, understanding, sharing, archiving and warehousing. Yeung & Hall [2007] adapted the metadata principles of Duval *et al* [2002] to provide the following principles for metadata standards, which are all catered for by ISO 19115 [2003] and its related standards, for example.

- **Modularity.**  
This is a key organising principle for the user environment, to accommodate the high degree of diversity of sources, contents and approaches to resource description, to facilitate combining metadata with different schemas, vocabularies or themes.
- **Namespaces.**  
These are the formal collections of terms managed according to a policy or algorithm that provide the mechanisms (eg: when using XML) to ensure global uniformity in the vocabulary used by a metadata standard. In Duval *et al* [2002], this is actually part of *modularity*, rather than being a separate principle. Effectively, a *namespace* is a form of a *controlled vocabulary*.
- **Extensibility.**  
This is to cater for types of metadata too specific or too detailed for general use, but (hopefully) without compromising the functionality of the base metadata schema. Yeung & Hall [2007] states that this allows for *profiles*, though strictly speaking, a profile is a *subset* of a standard (or group of standards), not a *superset*, with an extension implies.

## 5. Metadata

- **Granularity.**

This is the level of detail for the metadata, appropriate to a given application. Duval *et al* [2002] use the term *refinement* for this principle and describes two types: qualifiers making a metadata concept more specific (eg: the type of *creator*), and specifying the range for a metadata concept (eg: as a code list). In addition, *granularity* can indicate what items in the data are described by the metadata, ranging from a single field (*fine* granularity) through to a series of data sets (*coarse* granularity), and even beyond.

- **Multilingualism.**

This is to support cultural and linguistic adaptability (CLA) and not entrench one culture and language. Issues include the representation of dates, text directions, text sorting orders, name order and the cultural connotations of certain symbols. This is facilitated by encoding the metadata and by catering for multiple languages in one metadata set, as is done by ISO 19115 [2003], for example, and shown in Table 5.1 [Duval *et al* 2002; Yeung & Hall 2007].

Metadata and the standards for metadata can be generic or specific to a domain or discipline. McMahon [2015] suggests that a generic metadata standard (such as Dublin Core, see Section 5.8.8) will have relatively fewer metadata terms or elements, be static and semantically fluid, while a specific metadata standard will have more metadata terms, be extensible and semantically precise. He also suggests they will have different purposes (eg: cataloguing and identification for the generic standard, *vs* reproducibility, validation and retrieval for the disciplinary one) and users eg: librarians, digital archivists, repository managers and funders *vs* scientists, software developers and analysts) [McMahon 2015].

### 5.8.2 Some metadata standards

The International Organization for Standardization (ISO) has developed several standards for metadata — as at 1 October 2015, the ISO catalogue listed 49 International Standards and Technical Specifications with the word ‘metadata’ in their titles<sup>12</sup>. Some of these are abstract standards providing building blocks for creating the metadata standards that people would use, such as the six parts of ISO/IEC 11179 [2003–2005]. There are several that could be (and are) used by organisations producing geospatial data — fortunately, they share a common view of the nature of metadata, though they differ in the application areas they cover, their breadth, their depth, and their encoding.

The following are examples of metadata standards.

- ISO/IEC 11179, *Information technology — Metadata registries (MDR)*, a meta standard in 6 parts, see Section 5.8.3.
- ISO 19115:2003, *Geographic information — Metadata*, the ‘definitive’ metadata standard for geospatial data sets, see Section 5.8.4. The Open Geospatial Consortium,

<sup>12</sup>Up from ‘only’ 27 on 14 September 2011!

Inc (OGC) also uses ISO 19115 and its related standards for metadata in the OGC specifications.

- ISO 19115-1:2014, *Geographic information — Metadata — Part 1: Fundamentals*, the revision of ISO 19115, including geospatial services [ISO 19119Amd1 2008], see Section 5.8.5.
- ISO 19115-2:2009, *Geographic information — Metadata — Part 2: Extensions for imagery and gridded data*, see Section 5.8.6.
- ISO 19119:2005/ Amd 1:2008, *Geographic information — Services — Amendment 1, Extensions of the service metadata model*, caters for services that provide functionality through interfaces, such as searching, catalogue browsing, portrayal (displaying geospatial data on a device's screen), coordinate transformation, editing, generalization, querying or ordering products. Services can be chained together and aggregated, such as for a location-based service. ISO 19119 [2005] is not confined to services on the World Wide Web. ISO 19119 [2005] also catered for metadata for services, supporting the development of a service catalogue through the definition of service metadata. This was then improved by ISO 19119Amd1 [2008], and subsequently by ISO 19115-1 [2014], which superseded it.
- ISO/TS 19139:2007, *Geographic information — Metadata — XML schema implementation*, provides the implementation in XML of the comprehensive metadata profile of ISO 19115 [2003], defining the XML Schema. It provides a common specification for describing, validating and exchanging metadata. The XML was generated semi-automatically from the harmonized UML model of ISO 19115 [2003], though the conformance rules are not enforceable with XML Schema, because of limitations with XML Schema. For example, it is not possible to ensure that the metadata element *characterSet* has been documented if ISO/IEC 10646 [2011] is not being used.
- ISO 19115-3, *Geographic information — Metadata — XML schema implementation of metadata fundamentals*, provides the integrated implementation in XML of ISO 19115-1 [2014]; ISO 19115-2 [2009]; ISO 19119 [2005]. This makes it easier to combine and validate metadata for vector data with metadata for imagery and gridded data. ISO 19115-3 and ISO 19139-1 [2014] replace ISO 19139 [2007]; ISO 19139-2 [2012].
- ISO 19139-1, *Geographic information — Metadata — XML schema implementation*, provides encoding rules for implementing the metadata standards [ISO 19115 2003; ISO 19115-1 2014; ISO 19115-2 2009], updating those rules that were included in ISO 19139 [2007].
- ISO 19139-2, *Geographic information — Metadata — XML Schema Implementation — Part 2: Extensions for imagery and gridded data*, provides the implementation in XML of ISO 19115-2 [2009].
- SANS 1878-1:2005, *South African spatial metadata standard, Part 1 — Core metadata profile*, is a profile (subset) of ISO 19115 [2003] for South Africa, that also includes South African examples. Hence, even though ISO 19115 was “overprinted” as SANS 19115, it is too comprehensive for most South African data custodians, so it is still necessary to have SANS 1878-1 as the local profile. Further, SANS 1878

## 5. Metadata

---

will probably be imposed on data custodians through the SDI Act [South Africa 2003].

- The African profile of ISO 19115 [2003], see Section 5.8.7.
- ISO 19131:2007, *Geographic information — Data product specifications*, see Section 5.5.
- ISO/IEC 15444-2:2004, *Information technology — JPEG 2000 image coding system — Part 2: Extensions*, specifies image metadata and how to store it within a JPEG 2000 image.
- ISO 15836:2009, *Information and documentation — The Dublin Core metadata element set*, a widely-used but high-level metadata standard, see Section 5.8.8.
- ISO/IEC 15938-5:2003, *Information technology — Multimedia content description interface — Part 5: Multimedia description schemes*, specifies a metadata system for describing multimedia content, particularly in the MPEG format.
- ISO/TS 17369:2005, *Statistical data and metadata exchange (SDMX)*, is for harvesting national statistics from the Web sites of national statistical agencies. Rather than being a general metadata standard, it is really just an encoding of transport ‘meta-data’, such as ISO/IEC 8211 [1994], or even ordinary XML.
- ISO/IEC 19503:2005, *Information technology — XML Metadata Interchange (XMI)*, for the interchange of metadata in distributed heterogeneous environments between application development life cycle tools, especially UML and the Meta Object Facility (MOF) [ISO/IEC 19502 2005].
- ISO 15489-1:2001, *Information and documentation — Records management — Part 1: General*, provides business principles for the management of business records. It also provides a high-level framework for workflow for records management.
- ISO 23081-1:2006, *Information and documentation — Records management processes — Metadata for records — Part 1: Principles*, for archiving and records management in compliance with ISO 15489-1 [2001]. ISO 23081-1 does not define metadata elements but assesses the suitability for records management, of several existing sets of metadata elements.
- ISO 25577:2008, *Information and documentation — MarcXchange*, is an XML-based exchange format for bibliographic records and other metadata, particularly MARC (MAchine-Readable Cataloging).
- IEC 82045-2:2004, *Document management — Part 2: Metadata elements and information reference model*, is for the document management of technical drawings and the like.
- *Data Documentation Initiative (DDI), version 3.1*, is for describing data from the social, behavioural and economic sciences using XML, and was published in 2009. It aims at supporting the entire research data life cycle: conceptualization, collection, processing, distribution, discovery, analysis, re-purposing and archiving [DDI 2009].

- The *Content Standard for Digital Geospatial Metadata (CSDGM)*, version 2.0, was published by the American Federal Geographic Data Committee (FGDC) in 1998 and was a major source for ISO 19115 [2003]. It is being phased out in favour of the *North American Profile (NAP)* of ISO 19115 [INCITS 2009], but there is obviously a lot of legacy metadata in CSDGM.
- IEEE 1484.12.1-2002, *Draft Standard for Learning Object Metadata*, developed by the Learning Technology Standards Committee of the IEEE, was published on 15 July 2002, but appears to have died because it is still only a draft and because the Web site for it has not been updated since 2005 [IEEE LTSC 2016]. This illustrates that even when a strong standards generating body is behind a standard, it can still fail for a variety of reasons.

Unfortunately, an organisation might need to support several different metadata standards, because the requirements are driven by external stakeholders and because they cover different domains. However, *crosswalks* (ie: mappings) or even translation programs have been developed to enable conversion between several different metadata standards, such as FGDC's crosswalk from CSDGM to ISO 19115 [FGDC 2006] and the Library of Congress' crosswalk from MARC to Dublin Core [Library of Congress 2008].

Fortunately, many of the commonly used metadata standards have largely free-text formats, which makes it easy to dump metadata into them (though difficult to process metadata in such formats). The key is ensuring that metadata is captured into an encoded metadata format (such as ISO 19115 [2003]; ISO 19115-1 [2014]), from which free text can be generated readily for the other formats. The other advantage of an encoded format is that it should be relatively easy to generate metadata in different languages.

### 5.8.3 ISO 11179, *Information technology — Metadata registries (MDR)*

ISO/IEC 11179 is a suite of six standards for developing metadata standards, ie: they are *meta standards*. The parts describe what a metadata registry should contain, the kind and quality of metadata required, and how metadata should be managed and administered. They do not describe metadata *per se*, and hence are *abstract* and not *concrete* standards. This can make them confusing and difficult to understand how to use. The ISO/IEC 11179 standards are for the developers of standards for metadata, not for those who record and use metadata [ISO/IEC 11179 2003–2005]. However, some have used ISO/IEC 11179 as their metadata standard, such as Statistics Canada, though with extensions [Johanis 2005]. The six parts of ISO/IEC 11179 are:

- ISO/IEC 11179-1:2004, *Information technology — Metadata registries (MDR) — Part 1: Framework*.
- ISO/IEC 11179-2:2005, *Information technology — Metadata registries (MDR) — Part 2: Classification*.
- ISO/IEC 11179-3:2003, *Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes*.

## 5. Metadata

---

- ISO/IEC 11179-4:2004, *Information technology — Metadata registries (MDR) — Part 4: Formulation of data definitions*.
- ISO/IEC 11179-5:2005, *Information technology — Metadata registries (MDR) — Part 5: Naming and identification principles*.
- ISO/IEC 11179-6:2005, *Information technology — Metadata registries (MDR) — Part 6: Registration*.

### 5.8.4 ISO 19115:2003, *Geographic information — Metadata*

ISO 19115 [2003] together with its significant corrigendum, ISO 19115:2003/Cor 1 2006, *Geographic information — Metadata — Corrigendum 1*, has been the most widely used standard from ISO/TC 211, with Google Scholar [Google 2016e] recording over 7600 publications<sup>13</sup> referencing the standard, for example. It is possibly the most accessible of the ISO 19100 suite of standards, generating the most comments on any of the standards in the suite when drafts were circulated for balloting and comments (962 comments on the first Committee Draft (CD), 1500 on the second CD and 676 on the third CD). The revision of ISO 19115<sup>14</sup> has now been published, as ISO 19115-1:2014, *Geographic information – Metadata, Part 1: Fundamentals* [ISO 19115-1 2014], with 58 references to it already on Google Scholar.

ISO 19115 provides a rich conceptual model for metadata, that is independent of any specific encoding scheme. It makes extensive use of code and enumerated lists, which facilitates automated processing of the metadata and supports multiple languages, for example: see Section 5.4.

As with the other standards from ISO/TC 211, ISO 19115 uses the Unified Modelling Language (UML) to document the technical content of the standard, which facilitates harmonization with the other standards from ISO/TC 211. It has a data dictionary for each Section conforming to ISO/IEC 11179-3 [2003], providing a version of the UML diagrams that is easier for humans to understand. The standard also includes rules for extensions and profiles; a comprehensive dataset metadata application profile; an implementable metadata profile (see ISO 19139 [2007]); guidance on extending, implementing and managing metadata; hierarchical levels of metadata (at the level of attribute, feature, data set, etc); and multilingual support for free text fields that caters for the language, country variant and character set used for the text.

### 5.8.5 ISO 19115-1:2014, *Geographic information — Metadata — Part 1: Fundamentals*

ISO 19115 [2003] is the revision of ISO 19115 [2003]; ISO 19115Cor1 [2006]; ISO 19119Amd1 [2008], based on eight years of experience with using ISO 19115 [2003], to incorporate

---

<sup>13</sup>As at 28 July 2015.

<sup>14</sup>The project to revise ISO 19115 fell under ISO/TC 211's Working Group 7, *Information Communities*, for which I am the Convenor.

service metadata and to cater for technological changes, such as the greater use of the Internet. The revised standard caters for information and resources that can have geographical extents, including maps, specimens, artifacts and services. It can also be used to describe information resources that do not have a geographical extent.

Some of the problematic issues with the revision are backwards compatibility and catering for all the legacy metadata, so ISO 19115-1 [2014] is independent of ISO 19115 [2003]; ISO 19115Cor1 [2006] and the UML for ISO 19115Cor1 [2006] will remain in ISO/TC 211's harmonized model. Further, no new mandatory elements were created, some definitions were broadened and where a definition needed to be changed, the metadata element was deleted and replaced by a new one. ISO 19115-1 [2014] is still independent of any specific encoding scheme.

#### **5.8.6 ISO 19115-2:2009, *Geographic information — Metadata — Part 2: Extensions for imagery and gridded data***

This standard is an extension to ISO 19115 [2003], defining the metadata for imagery and other gridded data, such as:

- Properties of the measuring equipment used to acquire data;
- Geometry of the measuring processes employed by the equipment;
- Production processes used to digitise raw data;
- Derivation of geographical information from raw measurements;
- Properties of the measuring system;
- Numerical methods and computational procedures used; and
- Collection and processing of imagery.

As with ISO 19115 [2003], ISO 19115-2 provides a conceptual model that is independent of any specific encoding scheme. ISO 19139-2 [2012] provides the implementation in XML of ISO 19115-2.

#### **5.8.7 African profile of ISO 19115**

This profile of ISO 19115 [2003] was compiled by the Standards Working Group of the CODIST Sub-committee for Geo-information (CODIST-Geo). CODIST is the Committee on Development Information, Science and Technology, established to inform the United Nations Economic Commission for Africa (UN ECA) on development challenges. The African profile is essentially the core profile of ISO 19115 with minor adjustments, as agreed on at a CODIST workshop. The core profile is nominally only one table of 22 metadata elements in ISO 19115 [2003, Table 3], but it is fully expanded in the African profile — taking up several pages. Unfortunately, the African profile has been languishing as a draft since 2007, because UN ECA did not have the funding available to complete the standard.

## 5. Metadata

---

### 5.8.8 ISO 15836:2003 *Information and documentation — The Dublin Core metadata element set*

ISO 15836 [2009] defines the *Dublin Core Metadata Element Set*, which has 15 “properties for resource description”, that is, broad and generic types of metadata, for describing resources across domains, particularly document-like objects. It is managed by the Dublin Core Metadata Initiative (DCMI) [DCMI 2016].

ISO 15836 [2009] is a basic standard which can be easily understood and implemented and as such is one of the best known metadata standards [Higgins nd]. Dublin Core focuses on data discovery. The properties are primarily free text, though DCMI manages many element refinements (qualifiers) for various implementations or application profiles, as well as vocabulary and syntax encoding schema. The extended set of DCMI Metadata Terms has been implemented in the Resource Description Framework (RDF).

However, there are those who are sceptical about Dublin Core, because “it provides too few elements to describe complex resources and because of its separation from content standards” [Beall 2006a]. The Dublin Core extensions (element refinements) are “created at the local level” (ie: within an application environment), which makes it difficult to map them to other application profiles. Unfortunately, while Dublin Core metadata can be encoded in Web pages, it tends to be ignored by search engines because such metadata has often been used to misrepresent Web pages, to make them appear earlier in search results [Beall 2006a].

## 5.9 The limitations of metadata

Of course, even when provided comprehensively, metadata does not cure a faulty data set. However, the bigger problem is securing metadata in the first place: “*current standards-based approaches to metadata require considerable human input, and are difficult to maintain up-to-date. Moreover, they primarily represent the perspective of the data producer on the quality and utility of the data, and do not allow for users to express their measures of fitness-for-purpose*” [Craglia *et al* 2008], see Section 6.2.

Further, “*while metadata 1.0 has always struggled with the apparent unwillingness of custodians and producers of geospatial data to invest the time and effort required to create effective metadata, the experience of VGI suggests that many users who have spent time struggling with the problems of accessing and using a data set would be willing to contribute their war stories to such a repository. It also raises an interesting research question regarding the tools that the spatial accuracy assessment community might be interested in developing to help users make and contribute their own assessments*” [Goodchild 2008b].

Effectively, a resource (a product, a data set or a service) should perhaps have some sort of social media presence, to which such *war stories* could be contributed — and accessed by other users wanting to assess the fitness for their use of the resource. Metadata would probably be helped by the greater participation by users in the assessment of metadata and in providing feedback, as is done with commercial services such as Amazon or eBay

[Craglia *et al* 2008] and customer reviews on the likes of Tripadvisor and Expedia [DG Kourie 2015, *pers comm*].

- Details of what previous users found to be interesting and useful in a resource could enhance searching and retrieval [Craglia *et al* 2008].
- A resource itself could have a *hashtag* (a demarcated keyword or concatenated phrase) on the likes of Twitter, so that one could harvest the comments from users on the data set or service, for assessing its fitness for use.
- A resource could have the likes of a Twitter account or a Facebook page, where users could post comments on the product (including ‘photos’ of their activities with the data set or service, that is, screen shots of it being used by the user), and where the product could inform its users about its status, such as: update just released, update planned, maintenance budget scrapped, etc.

Doctorow [2001] identified seven “straw-men of the *meta-utopia*”.

1. **People lie.**

This is shown clearly by all the spam one gets. This really becomes a problem when there is no sanction for making exaggerated claims about one’s product described by the metadata.

2. **People are lazy.**

This is undoubtedly the biggest problem with creating metadata, because of the burden of doing so — even though the metadata invariably actually helps its author in the long run. As Doctorow [2001] points out, it is even difficult to get people to name their files sensibly! On the other hand, I am aware of some organisations that generate metadata for hundreds of data sets daily that conform to ISO 19115 [2003], and Craglia *et al* [2008] provide the example of the related field of multimedia data exploitation, which embeds much metadata within the data themselves.

3. **People are stupid.**

The result is that people make elementary errors in their metadata, such as preventing discovery of their products because they are incorrectly labelled.

4. **Mission: Impossible — know thyself.**

This applies particularly to user-generated content, but one would hope that professionals would document the processes undertaken to generate their product! Doctorow [2001] points out that as humans are poor observers of their own behaviour, it creates a massive market for religions and therapists! It also helps the likes of marketers, entrepreneurs, advertisers, journalists and politicians [DG Kourie 2015, *pers comm*].

5. **Schemas aren’t neutral.**

This is actually about classification, where the structuring of the classification system can bias perceptions about the classes, as discussed in Section 2.4.5.

6. **Metrics influence results.**

This is similar to the previous point, though perhaps more explicitly about metadata than the classification scheme. Whatever elements get included in, or excluded

## 5. Metadata

from, the metadata will bias perceptions of the product being described. This applies particularly to the dimensions and sub-dimensions of quality that are measured and reported, see Sections 6.5 and 6.6.

### 7. There's more than one way to describe something.

Doctorow [2001] suggests that *“requiring everyone to use the same vocabulary to describe their material denudes the cognitive landscape, enforces homogeneity in ideas”*. On the other hand, having a standard terminology helps those new to the field [Lemmen *et al* 2011].

Nevertheless, Doctorow [2001] acknowledges that metadata can be useful, particularly when generated or harvested automatically, rather than being created by humans. In commenting on this piece by Doctorow, Swartz [2013] stated *“utopian fantasies of honest, complete, unbiased data about everything are obviously impossible. But who was trying for that anyway? The Web is rarely perfectly honest, complete, and unbiased — but it's still pretty damn useful. There's no reason making a Web for computers to use can't be the same way”*<sup>15</sup>.

## 5.10 Metadata vs searching

Doctorow [2001] is not the only metadata sceptic, of course. As he is a Google employee, Parsons [2013] is obviously biased, and I recall him ‘demonstrating’ why search was better than metadata at the FOSS4G 2008 Conference in Cape Town. At another conference [All Things Spatial 2009], Parsons is reported to have stated that the then SDIs separated the metadata from the data, as in the “classic concept of a library”: electronic or card catalogue in one place, and the shelves of books elsewhere. Parsons reportedly claimed that it would be better to follow the evolutionary approach of the Web in general, exploiting “unstructured text search capabilities delivered by Google and other search engines (dynamically indexed and heavily optimised for performance)”, and data formats that can be indexed by search engines, such as GeoRSS, GeoJSON and (obviously as it is a Google product!) KML [All Things Spatial 2009].

On the other hand, the librarian Beall [2006a]<sup>16</sup>, states that it is better to have the metadata exist separately from the objects described, as the metadata is a surrogate for, and pointer to, the objects (obviously, this needs persistent identifiers for the objects). Metadata embedded in an object might only be found after the object has been found and examined.

Beall [2006a] also explains why full-text searching does not work well for serious searching:

- The search precision is poor — many, if not most, results are false hits;
- The term under discussion is not actually used on the retrieved page, because it is actually in the metadata for the page but is treated by the search engine as being

<sup>15</sup>Please note that this monograph, subtitled “An Unfinished Work”, was published with minimal editing after Aaron Swartz’s death, so it has grammatical errors.

<sup>16</sup>Now famous for his work on predatory publishers [Beall 2012, 2014d], see Section 4.8.3.

part of the page;

- The results are sorted according to the search engine's algorithms (eg: according to the search-engine optimisation performed on the returned Web pages, or even to promote paying advertisers), which might conflict with what the user needs;
- The search might miss relevant material spelt differently, or in other languages;
- The search might miss synonyms, acronyms and abbreviations, such as *river* vs *stream*;
- The search term does not appear in the text but in a graphic, such as a company's logo;
- Because of spamming and dubious search-engine optimisation practices, search engines often ignore the metadata encoded in Web pages; and
- Generally, the search engine does not present the results in a fashion that makes it easy for the user to scan through and select the most likely candidates for their need [Beall 2006a].

Beall [2006a] prefers the results from a dedicated metadata search engine that uses rich, structured metadata with *authority control* in catalogues, such as MARC, and returns the results in a neat, complete, collocated, left-anchored index display (by subject, title, author, etc). Authority control means ensuring consistency in the metadata by using the same, persistent label for each object, such as town name, country name, person name, organisation name or book title, even when the name changes, and cross-linking between the labels when the objects are split or merged. Google is now incorporating metadata into its advanced searching functions, making them function more like online library catalogues [Beall 2010]. Further, indexing as a searching aid is also a form of metadata, describing aspects of the resource.

## 5.11 Metadata and linked open data

As discussed above in Section 3.5, the concept of *linked data* refers to publishing semantic connections between digital objects that (hopefully) have persistent and discoverable identifiers, such as URIs. This can be implemented using RDF triples: subject, predicate and value (object). The concept of *linked open data* (LOD) then refers to data that are not only linked semantically, but are also available through an open licence that allows them to be reused for free [Berners-Lee 2009]. LOD needs reusable formats (eg: RDF) and catalogues to enable effective opening, linking and reusing digital resources [Ding *et al* 2012]. Note that LOD is different from the *open world assumption*, namely that the truth of a statement is independent of whether or not it is known, or to put it another way, "the absence of a statement is not a statement of non-existence" [Hebeler *et al* 2009; Dunsire 2013]. That is, the Semantic Web can never have complete knowledge of everything, so while adding new information might contradict old information, it does not delete the old information. Hence, the inferences derived from LOD can change as new content is

## 5. Metadata

---

linked into the Semantic Web [Hebeler *et al* 2009]. Please note that the Semantic Web is based on the open world assumption.

Such semantic linking can be done through the data themselves (the digital objects or resources), the metadata describing the resources, and both. For example, linking through the metadata can connect resources from the same custodian, captured for the same purpose, or that are valid for the same time period (eg: for historical research). Various libraries have made their catalogue metadata available as LOD [Haslhofer & Isaac 2011]. Metadata can improve the usability of linked data, through facilitating the storing, preserving, accessing and comparing of linked data, catering for multiple languages (see Section 5.4), finding patterns and inconsistencies, and assessing the quality of the data (see Chapter 6). However, as well as the problems with metadata discussed above, there can be inconsistencies between metadata and the assumptions and interpretations of metadata [Zuiderwijk *et al* 2012].

The linking can also be at different levels of *granularity*, such as to a data series, a data set, an individual instance of a feature type, or even an attribute of the instance — and, obviously, to the metadata for each. As the granularity refers to the detail, it implicitly also refers to how ‘dumb’ (ie: limited) the resource or metadata might or might not be [Dunsire 2013]. Further, the linkages do not have to be peer linkages, so a feature instance in one data set could be linked to a different data set as a whole. For example, a temporary ocean buoy in one data set could be linked to a data set of the satellite imagery taken when the buoy was active. Note that with linked data, *granularity* can also refer to the aggregation or disaggregation of data, which can be done using the predicate and value in an RDF triple [Hobel & Frank 2014].

In terms of the relationship of user-generated content to metadata and LOD, Dunsire [2013] provides examples such as “OK for my kids” or “too childish for me”. He points to the problem of AAA: *anyone can say anything about anything*. This obviously applies to folksonomies (see Section 2.4.3) as well: “someone will say something about every thing — in every [linguistically] conceivable way” [Dunsire 2013].

### 5.12 VGI and metadata

As discussed in Chapter 7, the availability and quality of metadata for VGI is considered to be a problem with VGI, and hence affect the perception and use of repositories of VGI. Further, it can be difficult to validate metadata [Rak 2013]. Some consider the metadata for VGI to be inadequate for use in a spatial data infrastructure. Metadata is particularly important for VGI, because of the disparate range of VGI producers (some of whom are anonymous or unknown) and the limited institutional memory retained by most VGI repositories. Such metadata are “collected sporadically by countless individuals on a variety of devices under a myriad of circumstances” and hence are a challenge for any metadata standard [Tulloch 2014]. Nevertheless, while there have been many studies assessing the *quality of VGI* (eg: see Section 6.8.2), it appears that there have been few assessing the *quality of the metadata* for VGI (or for professionally-generated content, for that matter).

Bravo *et al* [2015] assessed the metadata provided for VGI in several repositories against the official Brazilian Geospatial Metadata Profile and found that there was a good match, that is, the metadata elements were generally populated. However, it appears that they did not assess the actual quality of the metadata, that is, if the metadata was any good. Olfat *et al* [2012] developed a spatial metadata automation framework, based on ISO 19115 [2003] and on extracting metadata elements from a Geography Markup Language (GML) [ISO 19136 2007] encoding of the data. Souza *et al* [2013] have developed a tool for capturing automatically some of the metadata for VGI and for allowing the user to populate other metadata elements and to edit the metadata. It includes a mechanism for ranking the contributors of VGI.

Rak [2013] suggests that tagging could be used to create some types of metadata for VGI, by combining similar tags into tag clouds, eg: correlating '500', '1:500' and '1/500' with one another and as the scale of the data. Essentially, this would be done using an ontology, see Section 2.4.4.

Standards for metadata are typically producer-centric and one approach to generating metadata for VGI would be to encourage the users of the VGI to provide their comments on, and experiences with, the VGI. These types of crowd-sourced user assessments are now common on the Web, such as for restaurants, hotels, books and films [Elwood *et al* 2012]. The user could interpret the available commentaries them self, or metadata could be synthesized from the comments, such as by using ontologies. Kalantari *et al* [2014] propose *Geospatial Metadata 2.0* to create metadata automatically in two ways: for their *explicit model*, exploiting folksonomies (see Section 2.4.3) because that uses tagging by users; and for their *implicit model*, monitoring search terms to create a database of keywords for metadata records. Giuliani *et al* [2016] propose a workflow for generating ISO 19115 [2003] metadata semi-automatically from a limited set contributed when a data set is added to a data publication server and harvested using Catalogue Services for the Web (CSW) [Nebert *et al* 2007a].

### 5.13 Summary and looking ahead

While this chapter provides the setting for subsequent chapters, it also makes important contributions as part of my research and this thesis. Drawing on Chapters 2 and 4, this chapter has provided details of the definition of metadata; aspects of metadata such as its relationship to quality; active and passive metadata; the benefits of encoding metadata; the relationship between a product specification and metadata; tools for capturing metadata; the different categories or types of metadata; principles for standards for metadata; selected standards; limitations of metadata; comparing searching to metadata; metadata and linked open data (LOD); and VGI and metadata.

Chapter 6 is closely coupled with this chapter and will now provide details of the quality of resources, particularly geospatial data. It will provide details of the different aspects of the quality of resources, the four stages for recognising the quality of a resource, GNSS errors, the dimensions of quality, challenges for the quality of VGI, the quality of three

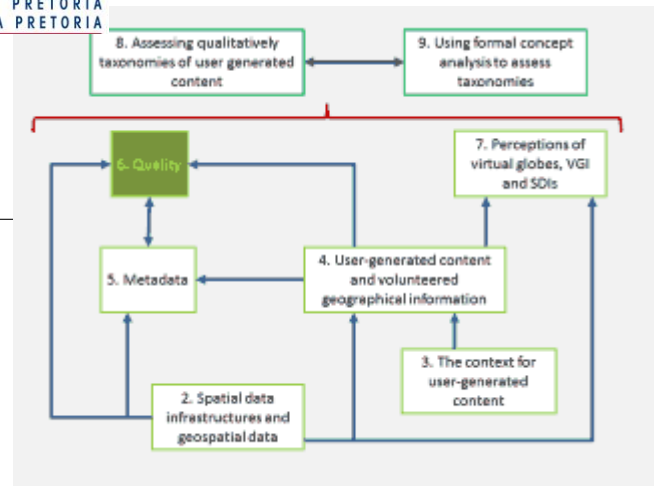
## 5. Metadata

---

VGI repositories, quality and classification, and standards for the quality of geospatial data.

\*\*\*\*





## Chapter 6

# Quality

### 6.1 Overview of the chapter

Chapter 2 discussed spatial data infrastructures and geospatial data, Chapter 4 discussed user-generated content and volunteered geographical information, and Chapter 5 discussed metadata. This chapter draws on all three to present a detailed discussion of **quality**, introduced in Section 2.7. Quality is closely coupled with *metadata*, and both are key characteristics of VGI repositories, which are assessed in Chapters 8 and 9. Further, the most common objections raised against VGI are perhaps the uncertainty over the quality of the VGI and of the documentation of the data (eg: Cooper *et al* [2010a], and see Chapter 7). Specifically, this chapter covers the following.

- Section 6.2 provides an overview of the different *aspects of the quality of resources*, particularly of geospatial data.
- Section 6.3 describes the *four stages* which I identified for *recognising the quality* of a resource.
- To illustrate these stages and give examples of quality problems, Section 6.4 presents some of the types of *errors* that one can get with *GNSS data*.
- The seven commonly-used *dimensions of quality* for geospatial data and their sub-dimensions are described in Section 6.5.
- Section 6.6 discusses *further perspectives* on the dimensions of quality and how they correlate with the dimensions discussed in Section 6.5.

- Section 6.7 explores the quality of volunteered geographical information against the dimensions of quality, and identifies several *challenges for the quality of VGI*.
- In Section 6.8, three VGI repositories are assessed against the quality dimensions (Section 6.5) and quality challenges (Section 6.7), namely the *Second South African Bird Atlas Project (SABAP2)* [Animal Demography Unit 2016b], *OpenStreetMap (OSM)* [OpenStreetMap 2016] and *Tracks4Africa (T4A)* [Tracks4Africa 2016]. A preliminary version of this analysis was published in Cooper *et al* [2012a].
- Section 6.9 considers *quality and classification* of geospatial data. Parts of Sections 6.8 and 6.9 were published as part of the paper Cooper *et al* [2011a].
- Section 6.10 discusses *standards* for the quality of geospatial data.

The major original contribution that I have made that is presented in this chapter is:

- I have identified the *four stages for the recognition of the quality* of a resource in general, see Section 6.3 and Figure 6.1.

Further, the key contributions that I have made that are presented in this chapter are as follows.

- I have presented the *dimensions and sub-dimensions of quality* cohesively, see Sections 6.5 and 6.6.
- I have identified some *challenges for VGI*, see Sections 6.7.2 to 6.7.4, which were included in Cooper *et al* [2011a].
- I have *assessed three VGI repositories* against the quality dimensions and quality challenges, see Section 6.8. A preliminary version of this analysis was published in Cooper *et al* [2012a].
- For a selection of repositories of VGI (which are discussed in Section 8.3), I have mapped the *responsibility for their specifications* against the *types of data* they contain, see Section 6.9.2 and Figure 6.9. This was also included in Cooper *et al* [2011a].

Hopefully, this work contributes towards a new and improved understanding of VGI, its quality and also its usability. Other authors have also contributed in this regard, for example, Elwood *et al* [2012] who aim “to frame the crucial dimensions of VGI for geography and geographers, with an eye toward identifying its potential in our field, as well as the most pressing research needed to realize this potential”.

Finally, this chapter raises some questions for further research.

1. How does one balance *maintaining the integrity of VGI* with making it *easy* for arbitrary producers of VGI to continue providing VGI, and to keep on improving quality?
2. In Section 6.2, how can one ensure that the *abstract model* used for the geospatial data will enhance quality?
3. In Section 6.8.4, how does one improve the *classification correctness, updating efficiency, completeness* (particularly ensuring consistent coverage across the whole of

## 6. Quality

---

the domain) and *metadata* for VGI?

4. Also in Section 6.8.4, how can one *involve VGI contributors* in the development of standards and protocols?

This Chapter draws on my involvement in various initiatives related to data quality over the last 30 years, such as the research of the International Cartographic Association's Commission on Spatial Data Quality; South Africa's Committee for Spatial Information (CSI) [Harvey *et al* 2012; Cooper 2013]; ISO/TC 69, *Applications of statistical methods*; and various projects at the CSIR, including for the then Department of Water Affairs [Olivier *et al* 1990]; for mapping the enumerator areas [Cooper *et al* 1996, 1997]; for guidelines for data content standards for Africa<sup>1</sup> [Cooper *et al* 2005]; and quality assurance of data for the South African Police Service (SAPS) on Innovation Fund projects [Elphinstone *et al* 1999].

Some of my PhD research was supported by a collaborative project with the University of Pretoria and the Wrocław University of Environmental and Life Sciences<sup>2</sup> [Cooper 2011c; Cooper *et al* 2011a]. Most significant was my work with Statistics South Africa (StatsSA)<sup>3</sup> [Cooper 2005], such as on the South African Statistical Quality Assessment Framework (SASQAF) [Statistics South Africa 2008].

### 6.2 Aspects of geospatial data quality

A resource (such as a geospatial data set) is only of value to a user when that user has information that will allow them to determine the quality of the resource. Unfortunately, there is a tendency for users to be unaware of the inaccuracies inherent in the data [Zargar & Devillers 2009], for example, being unaware of the issue of cartographic licence (generalization, aggregation, selective presentation, etc, see Section 2.9), being ignorant of the limitations of GNSS receivers (see Section 6.4), or being unable to distinguish between an error and a distortion, such as due to the particular map projection used. The core of the problem is that for different uses, significantly different levels of quality for the same data are necessary or acceptable [Cooper 1993], such as comprehensive *vs* timely *vs* accurate *vs* free [Ashley 2013].

Further, one needs to consider not just the quality of the data captured, but also the quality of the abstract model used for determining what geospatial data are actually included in one's database. Then, quality is maybe not just a technical issue, but also one where aesthetics could be an aspect of quality and quality an aspect of aesthetics<sup>4</sup>. Further, aspects of the quality of geospatial data are subjective, because they depend on the user of

---

<sup>1</sup>Initiated by the EROS Data Center of the United States Geological Survey (USGS/EROS) and EIS-Africa, with funding from the United States Agency for International Development (US AID) and the CSIR.

<sup>2</sup>Funded partially by a Joint Research Grant under the SA/Poland Agreement on Cooperation in Science and Technology.

<sup>3</sup>Two service level agreements funded by the CSIR and StatsSA, in terms of the Memorandum of Understanding covering the CSIR's participation in the National Statistics System (NSS).

<sup>4</sup>While not mentioned in his abstract, these issues were alluded to by Morita [2015] (a former ICA Vice-President) in his presentation at the recent International Cartographic Conference.

the data and the purpose and context in which they are used. Finally, most data quality metrics are really surrogates for the quality measures that really matter, such as truth [Ashley 2013].

The American National Committee for Digital Cartographic Data Standards (NCDCDS) did much to highlight quality issues for geospatial data, coining the terms *truth in labelling* and *fitness for use* [Moellering 1985].

- **Truth in labelling.**

The producer of the data must verify the quality of the data, and must furnish the prospective user with the details of the quality of the data. That is, the producer must be truthful in identifying the quality of the data.

- **Fitness for use.**

The prospective user must evaluate the quality of the information on data quality provided, and then decide whether or not the data are of a sufficient quality to be included in their database. That is, the recipient of the data must determine whether or not the data are fit for their use.

Perhaps the best-known family of ISO standards for quality is ISO 9000:2005, *Quality management systems — Fundamentals and vocabulary*, and its related standards. ISO 9000 [2005] defines *quality* as: “degree to which a set of inherent characteristics fulfils requirements”. As can be seen, this definition correlates well with the concepts of *truth in labelling* and *fitness for use*. The definition of *quality* that was used by the ISO 19100 suite of standards was similar: “totality of characteristics of a product that bear on its ability to satisfy stated and implied needs” [ISO 19101 2002], but it has now been replaced by the definition used in ISO 9000 [ISO 19101-1 2014].

The errors in data reflect the differences between the data and reality, and have two components.

- **Systematic error or bias.**

This is an error that has cascaded through the data set, such as when the location of the origin for a data set (eg: a trigonometric beacon) was recorded incorrectly; an aerial photograph was rectified incorrectly; an instrument (eg: a rain gauge or a GPS receiver) was used that had not been calibrated correctly; or an out-dated version of the classification scheme was used<sup>5</sup>.

- **Random error.**

This is an error that occurs only for one measurement, such as when coordinates are transposed (quite easy to do in South Africa, because the coordinate values for latitude and longitude are similar for a large part of the country); an instrument is not reset before taking a reading (eg: not emptying a rain gauge properly); or a typing error is made.

The combination of the *systematic* and *random errors* is the *total error*. A careless or inexperienced observer is more likely to make random errors than an experienced and careful

---

<sup>5</sup>This is a common problem with the 2nd South African Bird Atlas Project, for example, because some contributors use old bird lists and are unaware of the ‘recent’ splitting and lumping of species; see Figure 6.6.

## 6. Quality

observer, because the latter understands the whole data capture process and uses their experience to prevent errors. If there is structure to, or consistency in, the random errors, that would indicate a systematic error that could be corrected by recalibration (of equipment, datums, models, constants, etc) or by training, as appropriate.

Recently, Ormeling [2011] expressed the concern that technological cartography has left theoretical cartography too far behind. One example that he gave is that it is not yet possible to forecast the data quality of the result of combining different data sets together within a GIS.

### 6.3 Four stages of recognising the quality of a resource

As with metadata, the quality of a resource can be explicit or inferred. I have identified four stages for the recognition of the quality of a resource (data, product, service, process, transaction, operation, etc), though for many people, only one is generally considered; see Figure 6.1. Such a resource could be one of the repositories of VGI assessed in Chapters 8 and 9, which are described in Section 8.3. I have included *quality assurance* as one of the characteristics of the selected repositories in Table 8.2 and Section 8.3.3. Further, *quality* is one of the *issues* in the taxonomy of Budhathoki *et al* [2009] and is considered part of the *digital content policies* aspect of UGC by Wunsch-Vincent & Vickery [2007], but is not part of their taxonomy.

As discussed in Chapter 7, the quality of VGI and the documenting of the quality of VGI are considered to be problems with VGI, and hence affect the perception and use of repositories of VGI. Having a clearer understanding of the stages of recognising and documenting the quality of a resource — particularly over what gets lost between each stage — will help to improve the evaluation and reporting on quality, and the assessment of fitness for use. This is particularly the case for VGI, with the disparate range of producers (some unknown) and the limited institutional memory developed by most VGI repositories as they are so new and involve many amateurs (see the discussion on neo-geography in Section 4.7). In contrast, for example, the various South African national mapping series began in 1935 and have been evolving ever since with a documented history [Liebenberg 2014].

#### 6.3.1 Inherent quality

The *inherent quality* of any data (or product, service, process, transaction, operation or other resource) consists of the actual characteristics of the data that reveal how well the data represent the phenomena in the real or imaginary world that they are meant to represent. The inherent quality can be obvious or subtle; easy or difficult to comprehend or describe; crucial or practically irrelevant; and qualitative or quantitative, or both. Only a subset of the inherent quality is actually recognised and considered to be relevant, and hence passed on to subsequent phases. The inherent quality can exhibit *granularity*, that is, can be unique to a small part of the data (*fine granularity*) or generic to all the data (*coarse granularity*).

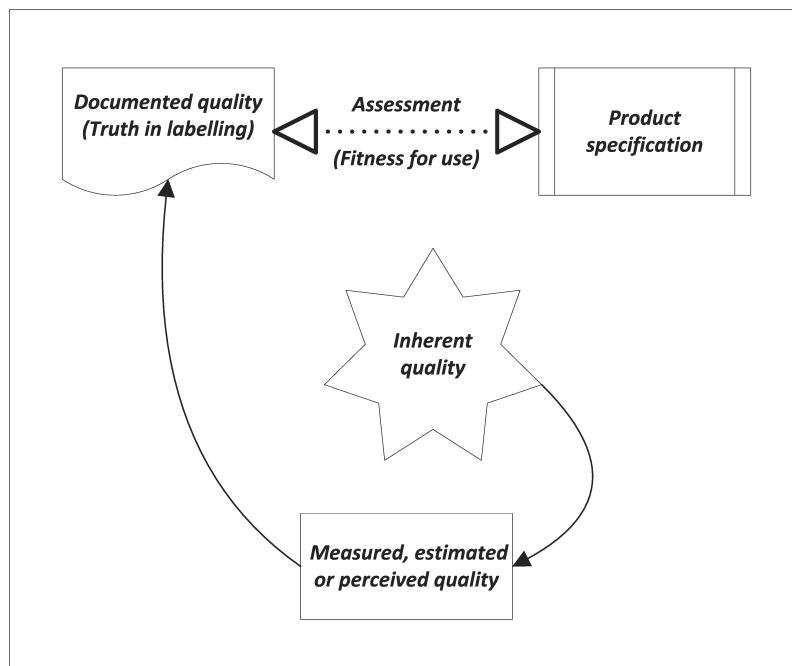


Figure 6.1: Stages of recognising data quality

### 6.3.2 Measured, estimated or perceived quality

Only some of the *inherent quality* is ever considered, because of factors such as cost, complexity or ignorance. The selected subset of the *inherent quality* can be considered in various ways, from rigorous measurement through estimations to perceptions, or a combination thereof. The approaches taken depend on the resources available and the expected costs and benefits of the alternatives. Hence, such assessment is at best only an approximation of the intrinsic quality: *measures, estimations or perceptions of quality* are always abstractions, always partial, and always just some of many possible ‘views’. The measured, estimated or perceived quality could apply to a dataset series, a dataset or a smaller grouping of data located physically within the dataset sharing common characteristics, so that its quality can be evaluated as a whole. Fortunately there is a standard for measuring the quality of geospatial data, ISO 19157 [2013]<sup>6</sup>, see Section 6.10.

### 6.3.3 Documented or reported quality

It is not enough just to *measure* or assess the *inherent quality*: it must also be *documented* to make it available to users or to be able to reuse the assessment later. This should be done in a systematic way, to make it as easy as possible for others to find and interpret the assessment of the inherent quality. It should also be possible for others to repeat the assessment and get similar results, in accordance with the science method. Hence, the

<sup>6</sup>ISO 19157 [2013] combines, updates and replaces ISO 19113 [2002], ISO 19114 [2003] and ISO 19138 [2006].

## 6. Quality

requirements for documenting the quality should dictate how the *inherent quality* is *measured, estimated* and/or *perceived*. The *documented quality* should adhere to the principle of *truth in labelling*, that is, care should be taken to document accurately what was assessed, rather than to interpret the results — that should be for the users to do themselves. For many, the documented quality of something is its quality, and the only phase generally considered. Fortunately, there are standards for documenting quality, such as ISO 19115-1 [2014] and ISO 19157 [2013]. It is likely that the *documented quality* will not include all the details of the *measured, estimated* and *perceived quality*, because of the need to condense the results into something more intelligible and useful for the average user, etc. Further, the *documented quality* is not necessarily made widely available.

### 6.3.4 Weighing up documented quality against a specification

As mentioned above in Section 5.5, a user should have a *specification* (their user requirement) for a data set (or product, service, process, transaction, operation or other resource), against which the *documented quality* can be weighed. The extent to which the *documented quality* matches the *specification* will indicate the *fitness for use* of the data, that is, how useful it would be for the user to use that data for their application. For example, ISO 19131:2007, *Geographic information — Data product specifications*, uses the metadata elements in ISO 19115 [2003] (and only them) to describe a specification. This then allows a direct comparison between the specification and the metadata of candidate data sets and facilitates semi-automated matching<sup>7</sup>, see Section 5.5.

### 6.3.5 ISO 19157 and the four stages

Figure 6.2 shows the conceptual model of quality for geographic data, from ISO 19157 [2013]. With reference to Figure 6.1 and the discussion above, this conceptual model does not include the *inherent quality* stage.

The *data quality scope* identifies for what resource the quality is evaluated: data set series, data set, a subset determined by type and/or extent, an individual feature instance or attribute value, etc. Each such set is termed a *data quality element*. The *documented* or *reported quality* stage is provided by the *standalone quality report* and/or the *metadata* ISO 19115, each using one or more of the *data quality elements* and related *data quality results*. Each *data quality element* should use only one *data quality measure* (the *dimensions of quality*, see Section 6.5), determined through the *data quality evaluation* (direct or indirect; by census or sample; wholly internal or using external sources, etc) to produce the *data quality results* (quantitative, conformance, descriptive and/or coverage results). The *data quality measure* and *data quality evaluation* together provide the *measured, estimated* or *perceived quality* stage. *Metaquality* describes how good the determining and reporting of the quality actually was, see Section 6.6.2.

*Weighing up documented quality against a specification* is out of the scope of ISO 19157 [2013], as it is dealt with by using a product specification based on ISO 19131 [2007], together

<sup>7</sup>The matching can be automated for those elements that are encoded, see Section 5.4.

## 6. Quality

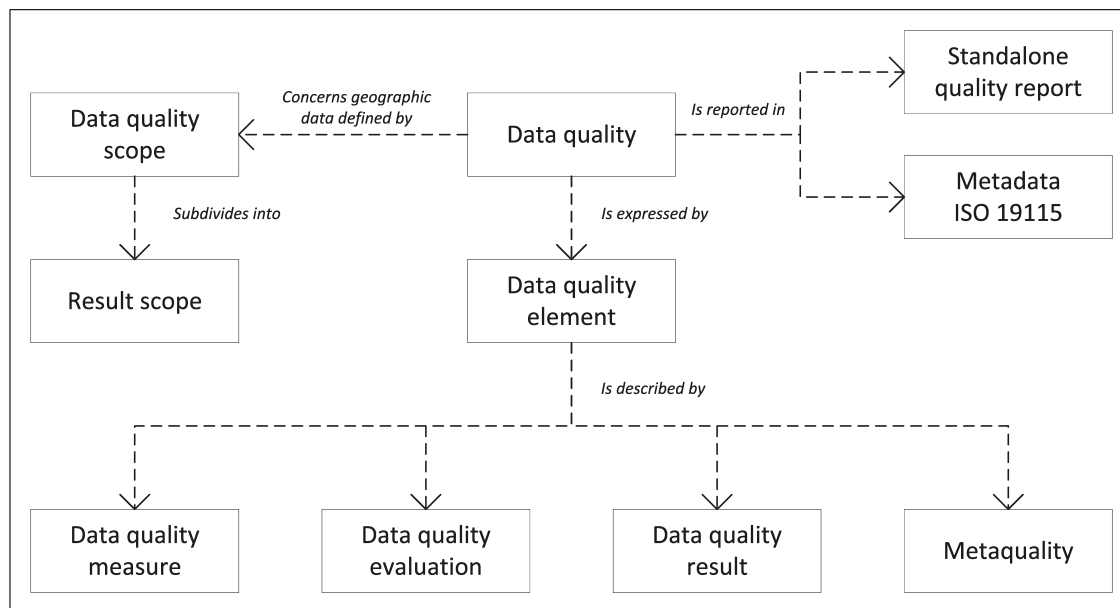
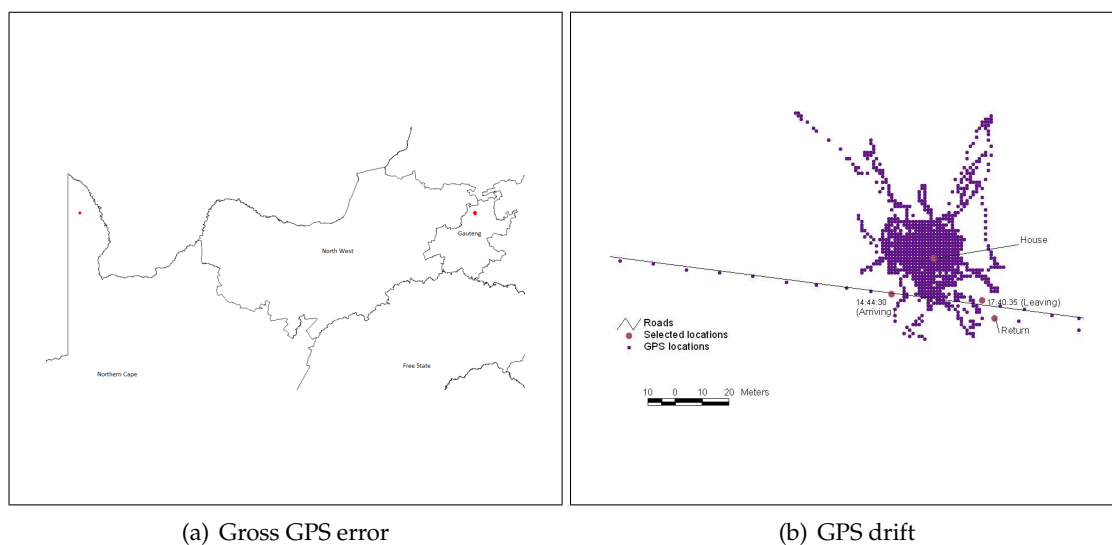


Figure 6.2: Conceptual model of quality for geographic data [ISO 19157 2013].

with the data quality from ISO 19157 [2013] and the metadata from ISO 19115 [2003] or ISO 19115-1 [2014], for the candidate resource. Nevertheless, ISO 19157 [2013] allows the producer to provide *usability elements*, which are essentially documented pre-assessments of the *fitness for use* of the resource for specific applications.

### 6.4 GNSS errors



(a) Gross GPS error

(b) GPS drift

Figure 6.3: GPS positional accuracy issues

## 6. Quality

---

To illustrate the four stages of recognising the quality of a resource, as introduced above in Section 6.3, consider the *positional accuracy* (as described in Section 6.5) of a consumer-grade GNSS receiver, such as is found in a PDA or cell phone. These are used widely for capturing VGI, as is shown in Table 8.3. Enormous faith is now placed in using GNSS for navigation, such as for maritime navigation, yet there are concerns over the vulnerability of GNSS and the lack of awareness of the risks by the relevant political and administrative leaders [Gutierrez 2014; Theunissen 2014]. It is also probably the case that few amateur users think of querying the positional accuracy of their GNSS receiver(s) — indeed, even professionals can be blasé about GNSS reliability [Gutierrez 2014]. However, the following issues do occur (see Figure 6.3).

- **Gross GNSS error.**

Figure 6.3(a) shows data that CyberTracker [2016] recorded from the GPS in my PDA on 25 October 2009, between 06:23:12 and 06:54:39, at the Rooiwal Sewerage Works north of Pretoria, whilst I was atlasing for SABAP2. As can be seen, the GPS recorded one position (at 06:33:02) as being in the Northern Cape: this was the 12th record in the sequence of 21 records. Table 6.1 shows the offending record and its predecessor and successor. Clearly, it would appear that the GPS or PDA changed an ‘8’ into a ‘0’. While this is the worst positional error I have encountered with this GPS, it is not the only one.

- **GNSS drift.**

Figure 6.3(b) shows the data recorded by a GPS device with a quality aerial mounted on a vehicle that was stationary for about 2 hours 55 minutes. The drift ranges up to 50m from the true position. These data were captured as part of the CSIR project, *GenDySI* [Cooper *et al* 2009a, 2010d; Schmitz & Cooper 2011].

- **Environmental influences.**

GNSS accuracy is affected by the ionosphere (affecting the phase and strength of the signal) and the troposphere (refraction affecting the signal path length). Closer to the receiver, the satellite signals can be reflected off surfaces such as high-rise buildings or cliffs, causing multi-path interference (or canyoning), or obstructed by objects such as trees and buildings (shadowing) [Van Niekerk & Combrinck 2012].

- **Selective availability.**

GPS was introduced by the American military and is controlled by them, initially with *selective availability* switched on, which limited the accuracy of all but military-grade GPS receivers. However, selective availability was switched off during the 1990s, though occasionally, it is switched on (globally or on a regional basis), such as during war time or for military exercises. For example, during October 2011, for Exercise Joint Warrior, the disruption to GPS reception off western Scotland was such that fishing and other civilian vessels were unable to navigate [McKenzie 2011]. The same applies to other GNSS services.

- **Jamming and spoofing.**

As the GNSS signal is relatively weak, it can also be *jammed* easily, intentionally or unintentionally [Van Niekerk & Combrinck 2012; Nighswander *et al* 2012; Rutkin 2013; Gutierrez 2014; Theunissen 2014; Lisi 2015]. Already, some airports are sub-

jected to multiple attempted jamming attacks daily [Theunissen 2014], and North Korea has apparently repeatedly jammed GNSS signals in South Korea [Lisi 2015]. Spoofing differs from jamming in that instead of blocking the GNSS signal, it is replaced by a fake one, which has already been demonstrated with equipment costing less than US\$ 3 000 [Lisi 2015].

- **Other failures.**

As with other types of satellites, even if they exceed their planned lifetimes, GNSS satellites do fail. Further, solar flares and storms can swamp the transmissions from the satellites [Gutierrez 2014; Lisi 2015]. Operator error can also cause a GNSS to fail, as happened with GLONASS for 12 hours on 1 April 2014, when invalid data were apparently uploaded [Lisi 2015].

Table 6.1: Three GPS records in sequence showing a gross error

Date	Time	Latitude	Longitude	Bird Name
2009.10.25	06:32:00	-25.556623333333	28.24524	Widowbird Red-collared
2009.10.25	06:33:02	-25.555	20.2452016666667	Lapwing Blacksmith
2009.10.25	06:33:48	-25.556655	28.245165	Wagtail Cape

Considering just the *positional accuracy* of a GNSS receiver under these circumstances and the four stages of recognising data quality (see Section 6.3), the receiver will have certain positional accuracy characteristics: its *inherent quality*. Only if the receiver is calibrated or otherwise assessed, will it also have *measured quality*, and only if these results are made available, will it have *documented quality*. Only then, can the user *weigh up documented quality against their specification*, should they so choose to do.

## 6.5 Commonly used dimensions of quality for geospatial data

Most, if not all, data and information products produced by government departments and related agencies, such as a statistics agency, have a geographical (spatial) dimension, and hence could be considered to be geospatial. When considering the quality of geospatial data, naïve users often consider only the positional accuracy of the data. However, there is more than just this aspect to the quality of spatial data — the following dimensions of quality have been widely recognised and used for geospatial data for the last three decades [Moellering 1985; Cooper 1993; Gupta & Morrison 1995; ISO 19113 2002; Bolstad 2005; van Oort 2006; ISO 19157 2013], see Figure 6.4.

However, while the top level of six of these seven quality dimensions might be well known, their sub-dimensions are not so well known. In comparison, for example, ISO 19157 [2013] does not include *semantic accuracy*; *geometric fidelity* and other details within *positional accuracy*; *currency* or *updating efficiency*. *Lineage* is included in ISO 19115-1 [2014], rather than in ISO 19157 [2013]. Then, *Semantic accuracy* is not widely used at this stage, probably because it is not well understood: see Section 6.5.3.

These quality dimensions are used to assess VGI in Sections 6.7 and 6.8.

## 6. Quality

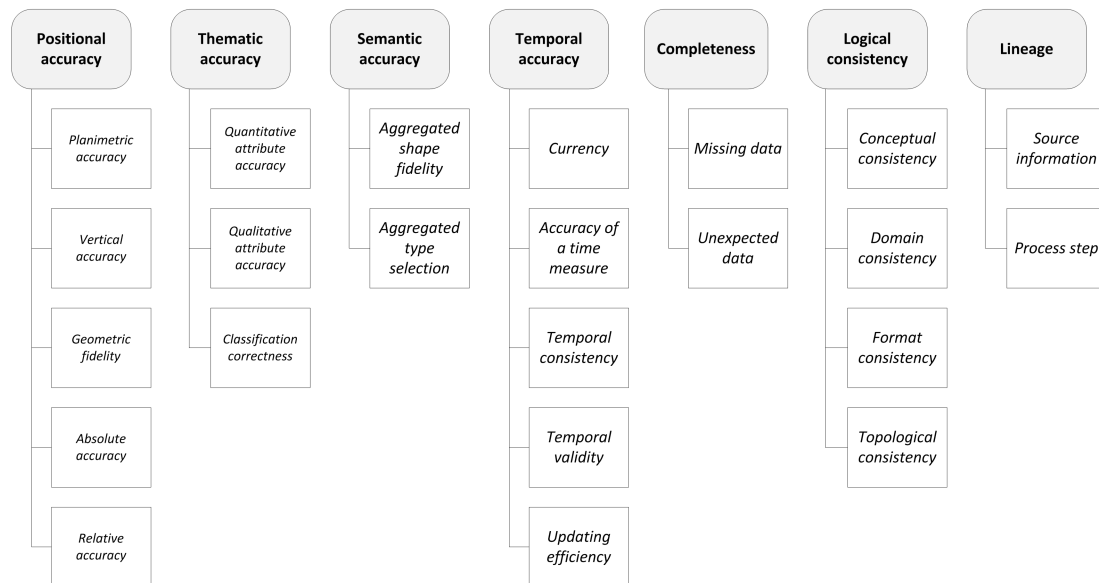


Figure 6.4: Dimensions of data quality

### 6.5.1 Positional or spatial accuracy

**Positional** or **spatial accuracy** is how closely the geometry of features in a digital data set correspond to the true locations of the related phenomena in the real or imaginary world. This applies to both vector and raster (or gridded) data. This has three components.

- **Planimetric accuracy.**

This is the geometric accuracy in the plane representing the surface of the Earth, defined with reference to a standard reference surface (typically, an ellipsoid), such as Clarke 1880 (modified), which was used for many decades in South Africa, and WGS 84, a global reference surface based on Global Navigation Satellite Systems<sup>8</sup> (GNSS), which became the standard reference surface for South Africa in 1999 [NGI 2016].

- **Vertical accuracy.**

This is the geometric accuracy above or below the plane of the Earth (height or depth), defined with reference to a vertical datum, most often a *geoid* (a complex model of the equipotential gravity surface at sea level), such as the Hartebeesthoek94 Datum for South Africa or the global Earth Gravity Model 2008 (EGM2008). Historically in South Africa, the vertical datum was the Land Levelling Datum (LLD), an approximation of a geoid based on the measurement of mean sea level at tide gauges [NGI 2016].

ISO 19157 [2013] does not have sub-elements for *planimetric* or *vertical accuracy*, but

<sup>8</sup>Such as the American Global Positioning System (GPS), the Russian GLObal NAvigation Satellite System (GLONASS), the Chinese Compass navigation system and the European Galileo positioning system.

it does have one for *gridded data position accuracy*. However, this is just a form of *planimetric accuracy*, for raster data alone, so it has not been included separately here.

- **Geometric fidelity.**

This is how closely shapes in the data set match the corresponding shapes in the real or imaginary world, and how closely the alignment of features in a data set match those of their corresponding phenomena in the real or imaginary world. This deals with aspects such as line of sight, orientation and relative positions [Ordnance Survey 2016]. Blana & Tsoulos [2015] refer to this as *shape similarity*.

Ideally, positional accuracy should be calculated on the basis of standard error or circular error, and be expressed in terms of a measure in the real world (such as metres), to make it independent of the scale of the data [Clarke *et al* 1987]. Planimetric and vertical accuracy can be measured in two ways.

- **Absolute or external accuracy.**

This is the closeness of the reported coordinate values for a position to the truth, or to values accepted as being the truth, in the real or imaginary world. The coordinate values need to be in some global coordinate reference system (eg: giving positions relative to the intersection of the Equator with the Greenwich Meridian), which is why it is the *external* accuracy. Otherwise, one can report only on their *relative* accuracy.

- **Relative or internal accuracy.**

This is the closeness of the offset between reported coordinate values to the geometric origin for the data set, and the corresponding offset in the real or imaginary world. Hence, relative accuracy considers both the accuracy of a coordinate value itself (given in a coordinate reference system covering only part of the earth — a *local coordinate system*) and the accuracy of the distances between recorded positions (ie: the consistency of positional data), which is why it is the *internal* accuracy. Relative accuracy is often used mainly for vertical coordinates, because they are much more difficult to measure accurately than planimetric coordinates — even with sophisticated technology, such as quality GNSS receivers.

It is not uncommon for a data set with a high relative accuracy to have a lower absolute accuracy, such as when the entire data set is shifted and/or rotated because of an error in the origin or framework used to position the data.

### 6.5.2 Attribute or thematic accuracy

**Attribute or thematic accuracy** is how closely the non-spatial attributes and the classification of features in a digital data set reflect the characteristics of the related phenomena in the real or imaginary world. Attribute values can be on nominal, ordinal, interval or ratio scales [Stevens 1946], can be free text, or can even multimedia. The classification assigns features to feature types or classes. This has three sub-dimensions.

- **Quantitative attribute accuracy.**

## 6. Quality

---

This is the closeness of the recorded quantitative data (ie: on the interval or ratio scales) to their values in the real or imaginary world. The error for a value can be quantified as a distance from the true value or, in the case of data on a ratio scale, also as a proportional error<sup>9</sup>.

- **Qualitative attribute correctness.**

This is the closeness of the recorded qualitative data (on the nominal or ordinal scales, or free text, multimedia, etc) to their values in the real or imaginary world. For data on a nominal scale, the correctness of one value can only be Boolean (true or false), because the nominal scale is merely an encoding of text strings. Even though an ordinal scale introduces a ranking of the values, one can also only state whether the value is correct or not: the problem is that the intervals in the ranking are arbitrary, so one cannot pronounce on the closeness of two values in an ordinal scale, whether they be neighbours on the scale or separated by many other values. Obviously, though, one can record aggregated accuracy rates for both nominal and ordinal data. For free text and multimedia, one can also report on how closely the data represent the characteristics of the related phenomenon in the real or imaginary world.

- **Classification correctness.**

This is the appropriateness of the classification of each feature in the data set for the phenomena in the real or imaginary world that they represent. If the classification is in the form of a linear list, then it is the same as attribute data on a nominal scale and the correctness of the classification for each feature can only be Boolean. If the classification has a hierarchy, then one can have partial matches, where a parent class is correct but the child feature type used is not. Again, because the ‘intervals’ in the classification are arbitrary, one cannot pronounce on the closeness of two feature types in a classification scheme, whether they be siblings with the same parent class or ‘separated’ by many other feature types. Again, though, one can record aggregated accuracy rates for the classification.

### 6.5.3 Semantic accuracy

**Semantic accuracy** is a measure that links the way in which the object is captured and represented in the database to its meaning and the way in which it should be interpreted. The difference between *semantic accuracy* and *thematic accuracy* is that the former deals with the transformation of data (particularly for generalization) and the appropriateness of the resulting classification and geometry (particularly shapes), rather than the correctness of the recording of the classes and attributes [Haunert & Sester 2008]. It would appear that the interdependence of *semantic* and *thematic accuracy* has not yet been assessed. *Semantic accuracy* has also not been used much to date (it has not been included in ISO 19157 [2013], for example), perhaps because it is considered to be esoteric or because few understand it. Nevertheless, the following could be sub-dimensions of *semantic accuracy*.

---

<sup>9</sup>The difference can be shown with temperatures in Centigrade (an interval scale) and Kelvin (a ratio scale because it has a true zero): 1°C is not 10% of 10°C, but is 9°C away from 10°C. On the other hand, 1 K is 10% of 10 K and is 9 K away from 10 K.

- **Aggregated shape fidelity.**

This is not just how closely an aggregated shape in the data set matches the corresponding aggregation in the real or imaginary world (as with *geometric fidelity*), but also how appropriate the aggregated shape is for that feature type. The former should be easy to measure but the latter requires interpretation and depends on context.

- **Aggregated type selection.**

This is the appropriateness of the feature type and attribute values selected for the aggregated feature. Hence, it is similar to the *attribute accuracy* for a single feature: the *classification correctness* of the aggregated feature type and the *quantitative attribute accuracy* and *qualitative attribute correctness* of the aggregated attributes.

#### 6.5.4 Temporal accuracy or quality

**Temporal accuracy or quality** is how closely the temporal data (time related attributes and relationships) in a digital data set correspond to the true values for the related phenomena in the real or imaginary world. As the measurement of time is on an interval scale (there is no true zero), the error for a value can be quantified as the difference in time from the true value, but not as a proportional error. One can also record aggregated accuracy rates for each for the five sub-dimensions of temporal accuracy.

The temporal component of geospatial data is generally considered from the ‘modernist’, ‘industrial’, ‘monochronic’ or clock-oriented perspective, which is context-free (*Sed fugit interea, fugit irreparabile tempus* [Virgil (Publius Vergilius Maro) 29 BCE]<sup>10</sup>), standardized and invariant; emphasises the likes of schedules (planned, established and maintained); and divides time into discrete, quantifiable blocks. This can create tensions and misunderstandings when applied to the other ontologies or constructs for time, be they ‘polychronic’, task-oriented, or based on the rhythms of seasons, weather, tides or animals [Bidwell *et al* 2013]. Such other paradigms might need additional sub-dimensions of temporal accuracy or quality.

- **Currency.**

This documents the time period(s) or time span for which the data are valid, which is not necessarily the present. It is critical to realize that the most recent data set is not necessarily the best, as it might be of a lower resolution or otherwise inferior. Further, historical data are needed for time series, archaeological, historical or other purposes.

- **Accuracy of a time measurement.**

This is the correctness of the temporal references of data elements (ie: that a date or time in the data is correct, within the specified tolerance).

- **Temporal consistency.**

This is the correctness of the ordered events or sequences, that is, whether or not

<sup>10</sup>“But meanwhile it is flying, irretrievable time is flying”. This ancient quote from Virgil shows that clock-oriented time is actually pre-modern and pre-industrial!

## 6. Quality

a sequence of events reported in the data is given in the order in which the events actually occurred. The error for a value can be quantified as the difference in time from its likely position in the sequence, assuming the other values are correct.

- **Temporal validity.**

This is the appropriateness of the data for the time span of the data set. Examples of data failing this test would be showing a railway line in the data set after the tracks have been removed, and showing erven in a data set before the township existed — sometimes, data sets are produced in anticipation of what will be there when the data set is published (particularly data sets with long lead times, such as maps in a national mapping series), but then the proposed development does not take place. The minimum bound for the error for a value can be quantified as the difference in time it is from the closest end of the time span when it would have been valid. While *currency* indicates when the data are meant to be valid, *temporal validity* indicates the extent to which this is the case. Unfortunately, this difference is lost on some, see Van Oort [2006, p 17] for an example.

- **Updating efficiency.**

This is the extent to which the *rate of update* for the data is suitable for the *rate of change* for the corresponding features and/or attributes in the real world. The gap between these is the *temporal lapse*, which can be given as a difference, but not as a proportion. When there have been sufficient updates to produce meaningful results, the trend of, or aggregated rates for, *updating efficiency* can be given. *Updating efficiency* also includes the date (and time, if necessary) of the *last update*. The *temporal validity* can be determined from the *rate of change* and the *last update*, when they are assessed against the *currency*. *Updating efficiency* also includes the expectation users can have of the data producer publishing according to their schedule, that is, their *punctuality*.

If the *time measurement* has been consistent (though not necessarily accurate) throughout the data set, then the *temporal consistency* should be good. However, the *temporal validity* is independent of both.

### 6.5.5 Completeness

**Completeness** concerns how well the data set represents all that it is meant to represent, and only what it is meant to represent, that is, the presence or absence of data and data elements. One key problem with completeness is having a common understanding of what is meant to be in the data set — for example, whether or not a data set of roads includes tracks or footpaths. These errors can be assessed by considering the data set as a whole or by identifying individual errors. There are two sub-dimensions.

- **Missing data.**

This is when data are absent from a data set, that is, data that are expected to be in the data set, but are not. These are errors of *omission*.

- **Unexpected data.**

This is when excess data are present in a data set, that is, data that are in the data set but that should not be there. These are errors of *commission* and include duplicates. Unexpected data are not necessarily a bonus or useful — for example, when using a data set of roads for routing large delivery vehicles, but it includes tracks, footpaths and other types of ‘roads’ that one would not want the heavy vehicles to use.

### 6.5.6 Logical consistency

**Logical consistency** is the degree of adherence of the data to logical rules of data structure, attribution and relationships — that is, the presence, absence or frequency of inconsistent data, such as inappropriate attributes for a feature or mismatches across data set boundaries. These errors can be assessed by considering the data set as a whole or by identifying individual errors. There are four sub-dimensions.

- **Conceptual consistency.**  
This is the adherence to rules for the data set. Examples of errors include inappropriate attributes for a feature and inappropriate relationships between features (eg: including a tree in a road transport network).
- **Domain consistency.**  
This is the adherence of values to their value domains, including those where the domain has an absolute boundary (eg: finding negative values for the size of a population), or where the domain has an expected boundary (eg: the typical ages of fertility for a woman<sup>11</sup>).
- **Format consistency.**  
This is the degree to which data are stored in accordance with the physical structure of the data set. Examples of errors include missing or invalid delimiters, incorrect character set used, and mismatched tags (eg: for data in XML or GML).
- **Topological consistency.**  
This is the correctness of the explicitly encoded topological characteristics of a data set. This applies specifically to the spatial component of the data set. Examples of errors include sliver polygons, boundaries that criss-cross each other, rivers that criss-cross contour lines, and mismatches across map-sheet boundaries.

### 6.5.7 Lineage

**Lineage** concerns the history of the data, including the people and organisations responsible for each stage. To the extent known, it should recount the life cycle of a data set, from collection and acquisition, through compilation and derivation, to its present form. As discussed in Sections 2.7 and 5.3, lineage could be considered as a dimension of *metadata* rather than of *quality*. Lineage has two sub-dimensions.

---

<sup>11</sup>I recall finding in a draft demographic data set of the South African population, over 100 women who allegedly had their first child at the age of 49 or older, and proceeded to have 16 or more children!

## 6. Quality

---

- **Source information.**

This describes the parentage of the data set, that is, the data sets, documents, imagery, measurements, observations, models and analyses that were used to create the current data set.

- **Process step or history information.**

This is the record of the events, methods or transformations in the life of the data set, including those used to maintain the data set, whether continuous or periodic, and their lead times.

Typically in the lineage, there will be ‘chains’ of sources separated by process steps, with some process steps combining multiple sources and/or spawning multiple sources.

## 6.6 Further perspectives on the dimensions of quality

### 6.6.1 Applying these quality dimensions to other resources

Clearly, to varying extents these dimensions and sub-dimensions of quality for geospatial data can apply to other resources, such as products, services, processes, transactions, operations and so on. Most obviously, every resource will have a *lineage*, though not necessarily with *source information* specifically between each *process step*. A Web service for coordinate transformation could have *planimetric*, *vertical*, *absolute* and *relative accuracy*, and a product such as a paper map could inherit all the quality dimensions of the geospatial data from which it was created, for example.

### 6.6.2 Quality dimensions identified by Van Oort

Van Oort [2006] also gives the following dimensions, taken from the literature he reviewed — in this case, primarily the old European standard, ENV 12656:1998, *Geographic Information — Data description — Quality*, that was superseded by ISO 19115 [2003]; ISO 19157 [2013]. They have not been included separately in the list of the dimensions of data quality above, because they are either metadata or they are already parts of the above dimensions of quality.

- **Usage, purpose and constraints.**

These are actually different aspects of the *metadata* that are used for determining the fitness for use of the data set. Hence, they are not a dimension of quality.

- **Variation in quality.**

This indicates the extent to which the errors are systematic or random (see Section 6.2), and as such, applies to all the dimensions and sub-dimensions of quality individually, as well as to any combination of them that might be suitable. For example, Haunert & Sester [2008] use only *logical consistency* and *semantic accuracy* for their assessment of generalization.

- **Meta-quality.**

This is the merit of the information provided on data quality. As with the *variation in data quality*, this applies to all the dimensions and sub-dimensions of quality individually, as well as to any combination of them that might be suitable. ISO 19157 [2013] caters for three types of meta-quality:

- *confidence* in the trustworthiness of the results obtained;
- *representativity*, the degree to which the results represent the truth about the population; and
- *homogeneity*, the similarity of the results obtained for different parts of the data set.

- **Resolution.**

As acknowledged by Van Oort [2006], this is generally found as a part of the elements of data quality and metadata, but see the discussion below.

### 6.6.3 Resolution *vs* accuracy *vs* precision

Unfortunately, there is also a tendency to confuse the terms *accuracy*, *resolution* and *precision*, which can make it complicated for a less experienced user to assess the quality of their VGI:

- **Accuracy.**

This is the closeness of observations, computations or estimates to the true values, or to the values that are accepted as being true [Moellering 1985]. Higher accuracy therefore implies that a measurement is nearer the truth, with the truth being either absolute or relative. Accuracy is the final measure of the worth of the data [Clarke *et al* 1987].

- **Resolution.**

This is the smallest unit that can be detected. Resolution provides a limit to *precision* and *accuracy* [Moellering 1985]. Spectral resolution is the width of different bands of the electromagnetic spectrum in which a multi-scanner operates. The spatial resolution of digitizing equipment is the minimum distance that the equipment can detect between any two points, while the spatial resolution of a plotter or printer is the minimum distance between plotted points (eg: dots per inch, or DPI).

- **Precision.**

This is primarily, a statistical measure of repeatability. It is usually expressed as the variance or standard deviation of repeated measurements [International Cartographic Association 1980]. However, in computing, the term is also used for the number of bits allocated to a number, and hence the fraction that can be resolved between two numbers. This is generally known as *storage* or *digital precision* [Van Oort 2006].

Pfuhl & Biegler [2011] considered *accuracy*, *resolution* (though they term it *precision*!) and *precision* (though they term it *reliability*) against one another, in an experiment using expe-

## 6. Quality

---

rienced orienteering runners. They found that poor resolution significantly affected the ability of the runners to find controls.

### 6.6.4 Perspectives of statistical agencies

Several national statistical agencies have considered in detail the dimensions of quality for statistical data and products produced by statistical agencies, both demographic and economic. They are discussed briefly here as they often have a geospatial component. Statistics Canada identified several dimensions of quality [Statistics Canada 2003], which were first published in 1998 and which have been adopted and expanded by other national statistical agencies, through the Committee for the Coordination of Statistical Activities (CCSA), which was established by the United Nations Statistics Division, and its Conferences on Data Quality for International Organizations [SDMX 2009; United Nations Statistics Division 2016; Vahed 2005; Statistics South Africa 2008]. These quality dimensions are assessed with reference to the dimensions of quality discussed in Section 6.5 above.

#### 1. Relevance.

This concerns whether or not the data and products meet the real needs of their clients. Rather than being a dimension of quality, this is really the assessment of the *fitness for use* of the data or product against a specification — that is, the fourth stage of the recognition of quality, described in Section 6.3 above.

#### 2. Accuracy.

This concerns whether or not the data describe correctly the phenomena they are designed to measure. For many people, this is the alpha and omega of quality. This obviously includes the quality dimension *thematic accuracy*, and would include *spatial accuracy* when the spatial component is included in the data. It probably includes some aspects of *temporal accuracy* (eg: *accuracy of a time measurement* and *temporal consistency*), together with *timeliness* and *punctuality*, as discussed below. It probably does not include *semantic accuracy*, which is possibly covered by *comparability*.

#### 3. Timeliness.

This concerns whether or not the data and products are available when needed. As with *relevance*, this is partially about the assessment of *fitness for use*, but it also includes aspects of the quality dimension of *updating efficiency*. Timeliness depends on the type of data and the applications of the data.

#### 4. Punctuality.

This concerns whether or not the data and products are published on their target dates. This is also an aspect of the quality dimension of *updating efficiency*.

The difference between *timeliness* and *punctuality* is that the former reflects the utility of the data or product when eventually released, while the latter reflects the ability of the agency to adhere to its published schedules. A product can be *punctual* but not *timely*, if the target date is too generous. If the target date is too ambitious

(and hence, also creates false expectations), the agency is likely to fail on *punctuality*, though *punctuality* does not determine the *timeliness* of the release.

Unfortunately, *timeliness* and *punctuality* together provide a limited view of the temporal dimension of data or products, assuming that only the “latest and greatest” data are all that matter — with the assumption that one works only with data relevant for the present. For many applications, it is also crucial to know the time span for which the data or products are meant to be relevant (the *currency* aspect of the quality dimension *temporal accuracy*), and how accurate the data or products represent that time span (their *temporal validity*), with the accuracy presumably decaying in one or both directions. *Timeliness* and *punctuality* should be supplemented by these other sub-dimensions of *temporal accuracy*, as discussed above in Section 6.5.

**5. Accessibility.**

This is how easy is it to obtain data and products, and their metadata, from the agency. This is not a reflection of the quality of the actual data or product, but reflects the operational effectiveness of the agency. Hence, it is beyond the scope of the dimensions of quality for *data*, as discussed in Section 6.5. *Accessibility* is an aspect of the metadata category *organisation metadata*, see Section 5.7.

**6. Interpretability or clarity.**

This concerns the ready availability of supplementary information and metadata to help users understand the data and products. This is a reflection of the quality and availability of the metadata, but not of the data or product *per se*, and hence, also beyond the scope of Section 6.5. *Interpretability* is an aspect of several metadata categories, especially *definitional*, *procedural* and *operational metadata*: see Section 5.7.

**7. Comparability.**

This concerns the extent to which differences in the reported statistics reflect actual differences in the true values. External factors that can limit *comparability* include changes in the geographical areas for which the data are gathered, changes in the environment over time and different methodologies used. It is common for statistical data to be aggregated, both thematically and spatially, and for such aggregations to be compared across space and time. Hence, *comparability* could be similar to the quality dimension of *semantic accuracy*. *Comparability* could also be part of the *implicit metadata* category, see Section 5.7.

**8. Coherence.**

This concerns whether or not the different data sets and products are harmonized within a broad analytical and temporal framework. This is the same as the quality dimensions of *conceptual* and *domain consistency*.

**9. Integrity.**

This identifies the extent to which the data and products are free from political interference, and the extent to which they adhere to the required levels of objectivity, professionalism, transparency and ethical standards. This is not catered for directly by the dimensions of quality in Section 6.5, but cuts across all of them: poor *integrity* for data or products should be reflected in some or all of the quality dimensions. The producer can provide metadata to substantiate their claim of *integrity* (pre-

## 6. Quality

---

sumably, all producers would like their data and products to be deemed to have integrity!), but the assessment would have to be done by the user, perhaps with the assistance of third parties, such as an independent quality assurance assessor. *Integrity* should also be an aspect of *organisation metadata*, see Section 5.7.

### 10. Credibility.

This concerns the extent to which the users are confident about the validity of the data and products, given their perception of the agency. Clearly, this reflects on the “brand image” of the agency and issues such as the agency’s perceived objectivity, professionalism, transparency, scientific merit and lack of manipulation. This is very similar to *integrity* and probably, either could be considered to be a subset of the other. As with *integrity*, the producer can provide metadata to substantiate their claim of *credibility*, but the user needs to do the assessment.

### 11. Methodological soundness.

This concerns the extent to which international standards, guidelines, good practices, agreed practices, and dataset-specific practices have been followed. While parts of *methodological soundness* could be included in the *process step* aspect of the quality dimension of *lineage*, any assessment of the soundness of the methodology would have to be done by the user. *Methodological soundness* is also an aspect of the *procedural metadata* category, see Section 5.7.

As can be seen, there are overlaps between quality and metadata. Some of these dimensions (particularly accessibility, interpretability and integrity) do not describe the quality of the data or products *per se*, but describe how efficiently and effectively the national statistical agency functions. This is not to doubt the importance of these issues to the end user, but they should be recognised for what they are, not ‘hidden’ as what they are not (ie: as data quality).

This has been recognized partially in the Data Quality Assessment Framework (DQAF) from the International Monetary Fund (IMF) and national and domain-specific derivatives of DQAF, such as the South African Statistical Quality Assessment Framework (SASQAF). DQAF and the others have *prerequisites of quality* for institutional and organisational conditions that impact on data quality [IMF 2016; Statistics South Africa 2008]. SASQAF defines four levels for the certification of sustainable series of national statistics in South Africa, determined by assessing each of 69 indicators (sub-dimensions) of quality and combining the results.

**Level Four: Quality Statistics.** These meet all the SASQAF requirements and could be certified as *official*, in terms of the Statistics Act [South Africa 1999].

**Level Three: Acceptable Statistics.** These meet most, but not all, of the SASQAF requirements and are fit for use for their designed purpose, despite their limitations.

**Level Two: Questionable Statistics.** These meet few of the SASQAF requirements and very limited deductions can be made based on them.

**Level One: Poor Statistics.** These meet almost none of the SASQAF requirements and hence no deductions can be made from them [Statistics South Africa 2008].

To some extent, *lineage* will be captured in some of these dimensions of quality of the statistical agencies, but there is much to lineage that will not be captured by them. *Completeness* is surely critical for the data and products of statistical agencies, but seems to have been omitted from their dimensions. With the inclusion of *completeness*, these dimensions are essentially the same as the quality criteria for big data that were identified by Cai & Zhu [2015]. In terms of VGI, the dimensions of quality from the statistical agencies are obviously relevant, but as discussed above, they are already embedded in the categories of metadata (see Section 5.7) and the dimensions of quality, see Section 6.5.

## 6.7 Quality of volunteered geographical information

Again, as discussed in Chapter 7, the quality of VGI and the documenting of the quality of VGI are considered to be problems with VGI, and hence affect the perception and use of repositories of VGI. Some consider the quality of VGI to be inadequate for use in a spatial data infrastructure. However, it has already been shown that this does not apply to all VGI, albeit indirectly, such as by Haklay [2010]. Various studies of the quality of the VGI in OpenStreetMap, in particular, have been conducted around the world, such as Govender [2011]; Mooney *et al* [2010a]; Du Plooy [2012]; Camboim *et al* [2015], and see also Section 6.8.2 for a fuller discussion of OpenStreetMap.

Having a clearer understanding of the dimensions of the quality of geospatial data will help to improve the evaluation and reporting on quality, and the assessment of fitness for use. These quality dimensions apply equally well to VGI as to professionally-generated geospatial data. With the disparate range of VGI producers (some anonymous or unknown) and the limited institutional memory retained by most VGI repositories, especially as they are so new and involve many amateurs, these dimensions are particularly important for VGI — as is determining how to apply them to VGI. However, the challenge is to balance maintaining the integrity of VGI (inherent, determined and reported quality) with making it easy for arbitrary producers of VGI to continue providing VGI, to retain existing and attract new producers of VGI, and to keep on improving the quality of VGI.

### 6.7.1 VGI and the dimensions of quality

Conceptually, the issues affecting the quality of VGI should be the same as those for professionally generated geospatial information, but there are differences. Table 6.2 considers how VGI in general maps to the dimensions and sub-dimensions.

The ready availability of cheap and reasonably accurate GNSS receivers means that the positional accuracy of VGI recorded using such a receiver should generally be accurate enough for most purposes at a scale of 1:25 000<sup>12</sup>. Typical errors that are likely to occur with amateur use of a GNSS receiver are transposing coordinates (as mentioned above, quite easy to do in South Africa, because the coordinate values for latitude and longitude

<sup>12</sup>That is, where 1mm on the map represents 25m in the real world.

## 6. Quality

are similar for a large part of the country), using the incorrect reference surface, or not waiting until a the GNSS receiver has recalculated its position.

Table 6.2: VGI and the dimensions of quality

Dimensions	VGI quality
<b>Positional accuracy</b>	
Planimetric accuracy	Likely to be reasonably high for most purposes when a GNSS receiver has been used, and used correctly, but see Section 6.4 for typical GNSS problems.
Vertical accuracy	When gathered, likely to be moderate because when a GNSS receiver has been used, because vertical accuracy is lower and more susceptible to distortions than planimetric accuracy.
Geometric fidelity	Likely to vary widely, depending on the diligence of the contributor, their equipment and most importantly, their understanding of the actual geometry of the feature they are digitising.
Absolute accuracy	Likely to be high when a GNSS receiver has been used.
Relative accuracy	Likely to be high when a GNSS receiver has been used.
<b>Thematic accuracy</b>	
Quantitative attribute accuracy	Likely to vary widely, depending on the diligence of the contributor, their equipment and most importantly, their understanding of the attribute they are recording.
Qualitative attribute correctness	Based on my experience of VGI on the likes of GoogleEarth and SABAP2, likely to be highly variable, depending on the contributor.
Classification correctness	Based on my experience of VGI on the likes of GoogleEarth and SABAP2, likely to be highly variable, depending on the contributor.
<b>Semantic accuracy</b>	
Aggregated shape fidelity	Unlikely to be a significant issue with VGI, because generalization is only likely to be done by experts.
Aggregated type selection	Unlikely to be a significant issue with VGI, because generalization is only likely to be done by experts.
<b>Temporal accuracy</b>	
Currency	Likely to be good if considered as a single moment or for a short time (eg: the pentade for a field sheet for SABAP2, see Section 6.8.1), but unlikely to be so for most VGI, unless the <i>updating efficiency</i> is high.
Accuracy of a time measurement	Likely to be very high when a GNSS receiver has been used.
Temporal consistency	Likely to be very high when a GNSS receiver has been used.

*Continued on next page*

## 6. Quality

Dimensions	VGI quality
Temporal validity	As discussed in Section 6.5.4, this could be confused with <i>currency</i> . However, it is likely to be high for VGI captured in the field as that is more likely to represent what is actually there, but could be low for VGI captured from imagery.
Updating efficiency	Likely to be good in metropolitan areas, because of the availability of VGI contributors, but poor in remote areas. However, VGI producers are unlikely to have publishing schedules, and hence <i>punctuality</i> is likely to be poor in general.
<b>Completeness</b>	
Missing data	Likely to be good in metropolitan areas, because of the availability of VGI contributors, but poor in remote areas <sup>13</sup> .
Unexpected data	Likely to be variable, depending on the contributor's understanding of the specifications.
<b>Logical consistency</b>	
Conceptual consistency	Likely to be variable, depending on the contributor's understanding of the specifications.
Domain consistency	While this depends on the contributor's understanding of the specifications, it is likely to be high for a VGI repository with good tools for automated bounds checking and correlating fields with one another.
Format consistency	Likely to be variable, depending on the contributor's experience.
Topological consistency	Likely to be variable, depending on the contributor's understanding of the specifications and diligence, and on the spatial model used by the VTGI repository.
<b>Lineage</b>	
Source information	Likely to be fair, because much VGI will have a short lineage chain with one source, the raw data captured in the field.
Process step	Likely to be good, because much VGI will have a short lineage chain with one standardized and well-understood process step between the raw data and the submission to the VGI repository.

### 6.7.2 Quality challenges for VGI

Drawing on observations made with my colleagues about user-generated content, volunteered geographical information and data quality, we identified several challenges for

<sup>13</sup>for example, the hard-copy *Botswana Traveller's Map*, published by Tracks4Africa, uses the VGI contributed to them, and while sufficient to produce a detailed map of the country, has poor coverage for those areas where few tourists go [Tracks4Africa 2011]

## 6. Quality

assessing the quality of VGI, which were published in Cooper *et al* [2011a] and are included here.

One of the biggest challenges is that due to the nature of VGI, some of the sub-dimensions of quality cannot necessarily be assessed at the time of contribution. Aspects of the quality of geospatial data are subjective, because they *depend on the user, purpose and context* of the application. Therefore, the contributor cannot assess the quality of their contribution in isolation. Rather, the user should assess the quality based on their *intended purpose and context*, using the information provided by the contributor about the data, that is, the *metadata*. However, as discussed above in Chapter 5, *metadata* is still not readily available for many data sets, particularly for VGI.

A further complication is that in general, *users are not involved in the development of standards*, such as for assessing quality or documenting metadata. The result is that even if they are aware of the relevant standards, they do not necessarily “buy in” to the standards nor understand their context or utility. Additionally, in our experience, even GISc professionals can struggle to read a standard without some training, because of the formal requirements for a standard and the necessarily repetitive structure of the text — a standard is not a novel!

Not all aspects of data quality can be assessed quantitatively, and as discussed in Sections 6.5 and 6.6.4 above, there are *important dimensions* of quality that have *qualitative aspects* to them. While quantitative measures can be understood in many languages (eg: root mean square error for positional accuracy), qualitative assessment is language dependent (eg: a statement about what should be included in the data set, for assessing completeness).

VGI can also be *contributed anonymously*, as in the annotation on Google Earth [Google 2016a] of sites allegedly connected with the pirates of Somalia, that was contributed by “expedition” [2009] as a KML file, see Figure 6.5. It would be very difficult for most people to verify the claim by “expedition” that “the pirate boats used for attacks are readily differentiated as slightly longer and broader than the normal pointed boats”, yet such a categorisation could have serious implications if wrong and acted upon.

Further, this illustrates how *transient data* can be: the boats might be at sea when an updated image is loaded on Google Earth and the KML would then point to an empty beach. This is a complex problem known as *incremental updating and versioning* [Peled & Cooper 2004] and is discussed above in Section 2.8. The key issue is that base data sets (such as the imagery in Google Earth) are generally updated without regard to the value-added content (not only VGI) that has been built on the base data and is dependent on the base data for its location or relevance.

Unfortunately, in addition to “normal” errors, not all contributions of VGI are made altruistically or without *bias*. Contributions could be made to promote a particular political, religious or social agenda; out of malice (eg: to denigrate someone or some community); with criminal intent (eg: to manipulate asset prices); or simply out of mischief [Coleman *et al* 2009]. Such malevolence can be in both commission and omission. Whereas poor data are likely to be poorly documented, malicious data might well have detailed metadata, albeit fraudulent! Of course, these problems can also apply to official data,

## 6. Quality

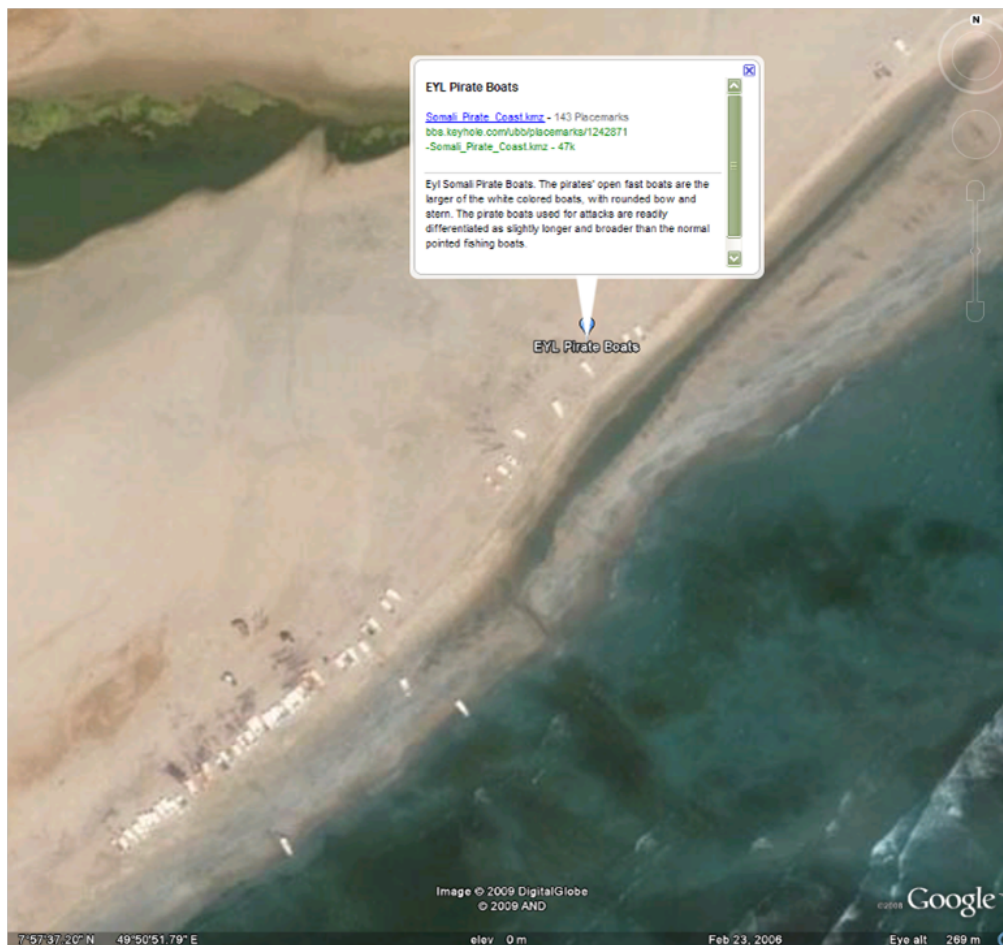


Figure 6.5: Alleged pirate boats on the beach at Eyl, Somalia [“expedition” 2009].

particularly from a repressive regime.

### 6.7.3 The risks

While mandated organisations (such as national mapping agencies) should produce data of higher quality than VGI, their mandates and priorities (eg: the need to provide national coverage or the need to support a specific national programme) might result in significant delays before they update data in certain areas. On the other hand, the public at large might be the best available source to keep local data up to date, such as verifying street names and addresses, or documenting changes when they happen and simultaneously submitting revision requests to the relevant agency (eg: as swisstopo allows [Guélat 2009]).

The risks of using poor quality VGI are primarily the same as the risks of using poor quality data from an official or commercial supplier — the source of the data will not affect the results of using the data. The key difference might be that an official agency

## 6. Quality

---

or commercial vendor could possibly be held legally accountable for their data, though in practice, this hardly ever happens because of disclaimers of liability (see Section 3.9 above).

### 6.7.4 Addressing these challenges

As with user-generated content in other contexts, such as Wikipedia [Wikimedia 2016], two key aspects of the quality assurance of VGI will be peer review, and peer pressure to adhere to norms, standards and the provision of metadata. The latter can be facilitated by the provision of automated tools for metadata capture and/or discovery. For example, the European Union's Joint Research Centre has recently released the European Open Source Metadata Editor (EUOSME) [European Union 2016], a Web application to create metadata in any of 22 European languages, that conforms to the requirements of INSPIRE, the European Union's SDI [European Parliament 2007], and of ISO 19115 [2003].

Another solution is to develop tools that automatically assess the quality of a specific VGI contribution, such as for logical consistency (e.g. valid attribute values), or against other data sources (as Tracks4Africa [2016] does). If these tools are deployed as Web services, they can then be used by more than one VGI repository.

As discussed in Section 6.9.1, aspects of quality could be part of a taxonomy of VGI.

## 6.8 Assessing the quality of several VGI repositories

Three repositories of VGI are assessed here: the Second South African Bird Atlas Project (SABAP2) [Animal Demography Unit 2016b], OpenStreetMap (OSM) [OpenStreetMap 2016] and Tracks4Africa (T4A) [Tracks4Africa 2016]. These assessments are done according to the dimensions and sub-dimensions of quality described in Section 6.5, and the quality challenges that are described in Section 6.7.2. These three repositories are also used with seven others to assess six taxonomies of user-generated content and one of citizen science. The assessments are done qualitatively in Chapter 8 and using formal concept analysis in Chapter 9. For SABAP2, please see also Section 8.3.2.7, for OpenStreetMap, please see also Section 8.3.2.3, and for Tracks4Africa, please see also Section 8.3.2.2.

Please note that a preliminary version of this analysis, for SABAP2 and OpenStreetMap, was published in Cooper *et al* [2012a].

### 6.8.1 Second South African Bird Atlas Project (SABAP2)

For the *Second South African Bird Atlas Project (SABAP2)*, begun in July 2007, volunteer bird watchers (and professional ornithologists) are gathering data according to a detailed, published protocol (recording bird distribution, observer effort and an index of

## 6. Quality

abundance) and submitting the data either directly to the SABAP2 Web site, or by sending them through the post [Animal Demography Unit 2016b; Wright 2011; Underhill *et al* 2012]. SABAP2 is managed by the Animal Demography Unit (ADU) at the University of Cape Town, in collaboration with the South African National Biodiversity Institute and BirdLife South Africa.

The data are gathered by geographical units, namely *pentads* (5' by 5'), and by temporal units, namely *pentades* (up to 5 days). As of 17 January 2016, 1853 observers have submitted over 145 200 field sheets, with 19 observers having submitted over 1000 field sheets each. While the raw data are not made available on the Web site, various processed data are available. SABAP2 builds on the First South African Bird Atlas Project (SABAP1, run from 1986 to 1997 [Harrison *et al* 1997, 2008; Underhill & Brooks 2014]), but with a more rigorous protocol with a finer spatial and temporal resolution that produces a better index of abundance. While increasing spatial resolution could be limited by the number of available observers [Robertson *et al* 2010], this has not been the case with SABAP2. The same platform and similar protocols are being used for other atlas projects, such as butterfly and reptile atlases [Animal Demography Unit 2016c,d]. SABAP2 is also recording data for other African countries, including Kenya and Tanzania.

SABAP2 now has sufficient data to produce distribution maps at a pentad scale for some bird species, particularly around the metropolitan areas [Underhill *et al* 2012; Underhill & Brooks 2014; Underhill *et al* 2014] — which have the higher concentrations of atlasers and hence more data, but which are also where the greatest environmental change is taking place. These distribution maps are showing alarming reductions in the numbers of some species (eg: Cape Cormorant, *Phalacrocorax capensis*), pleasing recoveries of others (eg: African Black Oystercatcher, *Haematopus moquini*, perhaps due to the ban of vehicles on beaches?) and range expansions of aliens (eg: Common Myna, *Acridotheres tristis*, and Common Starling, *Sturnus vulgaris*). SABAP1 and SABAP2 data are being used for environmental impact assessments (EIAs), such as for wind farms and power lines, but there has not been much assessment of the data quality (other than the limited amount of data for some areas of interest). Possible problems are that rare or endemic species could be over represented as observers might seek them out, but missed entirely in areas with limited records; varying observer capabilities; artifacts from the grid structure (pentads, for SABAP2) used for sampling [Robertson *et al* 1995, 2010]. Bonnevie [2011] cautions that biases can be introduced when comparing data from SABAP1 and SABAP2, due to their different spatial resolutions. Loftie-Eaton [2015] explains why the bias could result in higher or lower relative reporting rates, and discusses other problems with comparing SABAP1 and SABAP2, such as the better identification skills of birders now and better field tools, such as better field guides (eg: *Chamberlain's LBJs: The Definitive Guide to Southern Africa's Little Brown Jobs* [Peacock 2012]) and mobile electronic devices.

SABAP2 was selected for this analysis because I am familiar with it, having contributed 92 full protocol field sheets and 21 *ad hoc* protocol field sheets for 67 pentads, up to July 2015. For SABAP2, please see also Section 8.3.2.7.

- **Positional accuracy.**

Field sheets need to be located in the correct pentad (5' x 5', about 9km x 9km), ie: a matter of *relative accuracy*. The protocol is tolerant of errors greater than those of

## 6. Quality

consumer GNSS receivers, because it accepts observations of birds heard or seen close to the pentad boundary, and because the data are not meant for use at a large scale. For incidental observations, coordinates are preferred (ie: *absolute accuracy*), otherwise geographical identifiers with a narrative are accepted. *Vertical accuracy* and *geometric fidelity* are not relevant here.

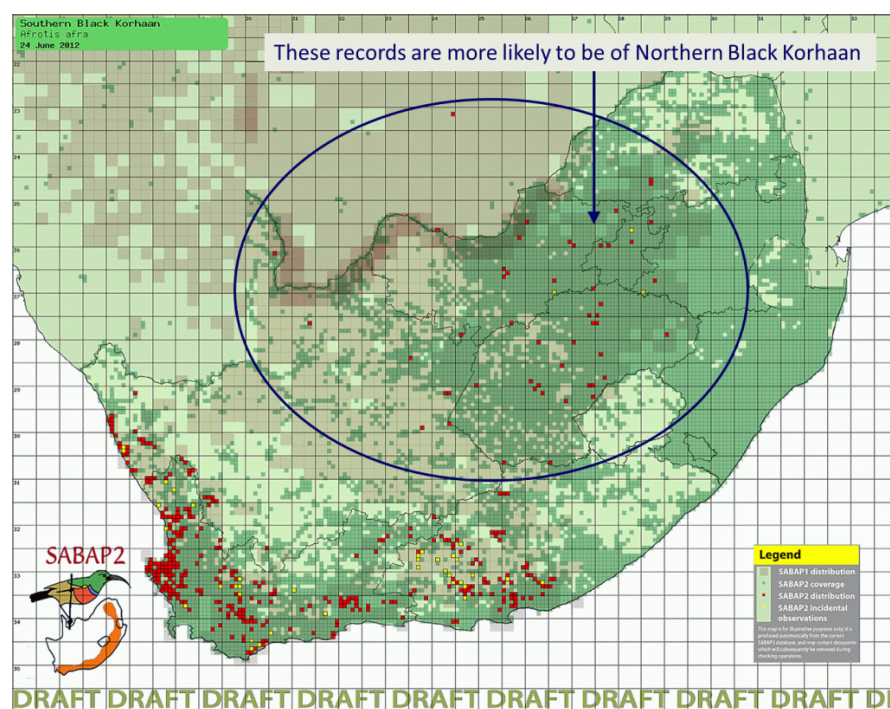


Figure 6.6: SABAP2 distribution of Southern Black Korhaan, as at 24 June 2012 [Animal Demography Unit 2016b].

- **Thematic accuracy.**

Probably the biggest problem (ie: *classification correctness*), as it depends on the identification skills of the observer, using jizz<sup>14</sup>, behaviour, plumage, calls, etc, and their knowledge of recent splits and lumpings in the taxonomy. However, the ready availability of multimedia electronic field guides (with pictures and calls) greatly aids in identifying species in the field [Robertson *et al* 2010], as I have found. Separating similar species can be particularly difficult where ranges are uncertain or overlap, or where species hybridize, as is the case with the Karoo Thrush, *Turdus smithi*, and Olive Thrush, *T. olivaceus*, species complex [Wilson *et al* 2009]. Of course, such SABAP2 data are particularly important for identifying the ranges of the species — particularly where there are confusing distribution maps and limited guidelines for visual separation of species, as is the case with Karoo and Olive Thrush [Wilson *et al* 2009].

The SABAP2 Data Management System can help by identifying when one is entering confusing species. Species out of range are identified automatically and the

<sup>14</sup>“The characteristic impression given by a particular species of animal or plant” [Oxford 2016].

observer is required to justify the observation, possibly submitting photos, videos or audio recordings. However, it uses the SABAP1 data to determine ranges. For example, Figure 6.6 shows the SABAP2 distribution for Southern Black Korhaan *Afrotis afra* as at 24 June 2012, but it probably includes records of Northern Black Korhaan *A. afraoides*, as indicated (the two species were split after SABAP1)<sup>15</sup>. The field sheets also cater for attributes, though they are not mandatory, so *qualitative attribute accuracy* (eg: number observed) and *qualitative attribute correctness* (eg: breeding status) can be relevant for some records.

- **Semantic accuracy.**

This is not an issue for the data capture (ie: for the VGI), but is for the analysis, which could involve aggregating, interpolating and smoothing data, by an expert. As the aggregation is not done by the observers, it is not relevant for the discussion on VGI quality here.

- **Temporal accuracy.**

For full-protocol field sheets, observations need to be for the correct hour (ie: *accuracy of a time measurement*) and provided in the order recorded (ie: *temporal consistency*). For *ad-hoc* field sheets and for incidental observations, observations need to be for the correct day. Many records are submitted months, or even years, after they were recorded, which could have accuracy problems, depending on the quality of the observer's record keeping. *Temporal validity* is not relevant here (records from before the project started will be trapped by the SABAP2 Data Management System), but *updating efficiency* is, as a measure of how often a pentad is re-surveyed.

- **Completeness.**

The more skilled the observer, the better the completeness for a field sheet, in terms of both *missing* species and *unexpected* species. Completeness could be estimated by comparison to SABAP1 and other data. Spatial coverage is not even, with over 72% of the pentads having been atlased at least once, and about 30% of the pentads being atlased each year. On the other hand, there are over 50 pentads (0.3%) that have been atlased 100 or more times. Hence, it is easy to determine exactly what the spatial completeness is for each pentad.

- **Logical consistency.**

The SABAP2 Data Management System presents some fields as drop-down boxes and does some consistency checks before a record can be submitted, and both work well to ensure consistency. Given the nature of the data being captured, *conceptual*, *domain* and *format consistency* will be trapped by the SABAP2 software, but not when an observer submits a record as a spreadsheet file or paper record (which are permitted). *Topological consistency* is not relevant here, because the observers do not submit geometric data.

- **Lineage.**

The observer for the submitted record is documented, together with any additional observers (ie: *source information*), so when analysing the data, it should be possible to determine the capabilities of individual observers in comparison to others. The

<sup>15</sup>Even worse, the ranges of Southern and Northern Black Korhaan actually overlap in the Karoo

## 6. Quality

---

*process steps* are not relevant here, as each field sheet or incidental observation is independent.

- **Dependence on purpose & context.**

A record does not have metadata indicating the abilities of the observer. However, since the record does identify the observer, one could use all their records to assess their abilities in comparison to other observers with records from the same or similar pentads and pentades. The metadata does provide indications of effort (hours spent observing and the number of new species per hour), so when analysing the data, it should be possible to determine the capabilities of individual observers in comparison to others.

- **Non-involvement in standards.**

There was little participation by potential contributors in developing the protocol for SABAP2. However, both the protocol and the software were modified early on, to accommodate suggestions from observers. Many workshops have been arranged all over the country to explain the protocol and the SABAP2 tool and experienced atlassers often take novices with them when atlassing.

- **Anonymous contributions.**

This is generally not possible, as an observer needs to register first, providing their contact details.

- **Bias.**

Potentially, a contributor could deliberately and malevolently exclude a species from a record to reduce its reporting rate, or include it to increase its reporting rate. This could be done to influence environmental impact assessments (EIAs), for example. Specifically, one could omit endangered species to help an EIA get accepted for a proposed development or one could include endangered species to make an area appear to be an environmental hot-spot, to prevent a development. These might not happen in SABAP2, but there is concern that unethical developers might manipulate such data. Similarly, contributors could include doubtful records to boost their standing amongst other atlassers.

Hence, while there is an emphasis on obtaining breadth of coverage in SABAP2 (getting records for as many pentads as possible), there is also an emphasis on obtaining depth of coverage (getting many records for each pentad), and both will reduce the vulnerability of any analysis to individual records. Some pentads might have lots of data, but from only one observer, eg: pentad 3015\_2555 has had one *ad hoc* and ten full protocol field sheets submitted between December 2007 and December 2014 inclusive, recording 112 species in total — but all of them were submitted by me, atlassing alone.

- **Qualitative aspects.**

Weather conditions are not recorded and are a big factor in the ease of identifying species in a pentad, but this could be obtained from other sources.

## 6.8.2 OpenStreetMap

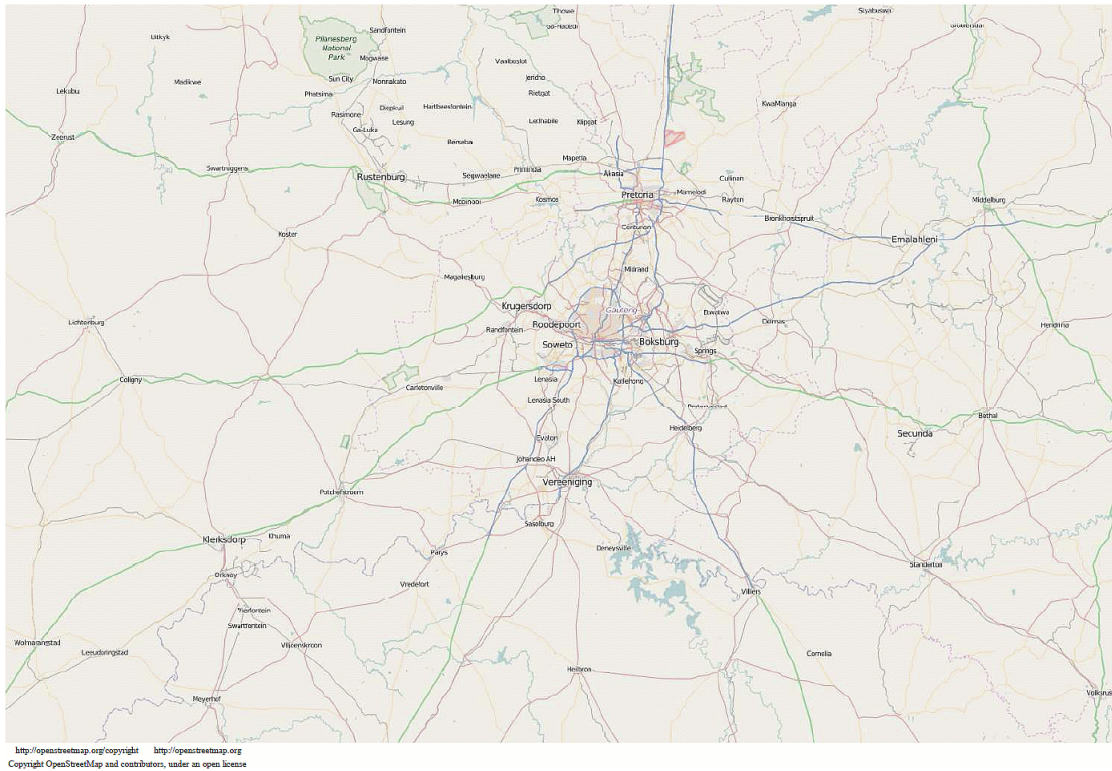


Figure 6.7: OpenStreetMap showing the Gauteng area [OpenStreetMap 2016].

*OpenStreetMap (OSM)* is a repository and a Web site providing a free, editable map of the whole world, initiated in 2004 as a repository of VGI, with Wikipedia [Wikimedia 2016] as its inspiration, see Figure 6.7. OpenStreetMap has a wide variety of tools for detecting possible errors (especially topological errors and missing tags) and procedures for publishing and correcting detected errors. OpenStreetMap data are sometimes more up-to-date and of a higher quality than commercial or official data sets [OpenStreetMap 2016]. OpenStreetMap data are widely used and are available through other Web sites. OpenStreetMap also contains much data contributed by official mapping agencies (including South Africa's national mapping agency, the Chief Directorate: National Geo-Spatial Information), but only VGI is considered here.

Several studies have been conducted to assess the quality of the data in OpenStreetMap in various ways, such as Ather [2009]; Kounadi [2009]; Haklay [2010]; Girres & Touya [2010]; Mooney *et al* [2010b]; Zielstra & Zipf [2010]; Borba *et al* [2015]; Camboim *et al* [2015], and in South Africa, by Govender [2011]; Siebritz *et al* [2012]; Du Plooy [2012]; Hankel [2012]. In collaboration with aid agencies, commercial satellite data providers and other organisations, OpenStreetMap made a significant contribution to mapping Port au Prince and other parts of Haiti for relief operations after the earthquake there on 14 January 2010, for example [Ball 2010a; Meier 2012; Zook *et al* 2012].

While OSM has over one million registered members who contribute data, the bulk of the

## 6. Quality

---

contributions or edits are done by a small proportion of the members, termed the “senior mappers” by [Mooney & Corcoran 2014]. This correlates with the *long tail* exploited by many Web services [Anderson 2004], as discussed in Section 3.4.12. It appears that these senior mappers work primarily on their own [Mooney & Corcoran 2014; Du Plooy 2012].

OpenStreetMap was selected for this analysis as it is probably the best-known repository of topo-cadastral VGI. For OpenStreetMap, please see also Section 8.3.2.3.

- **Positional accuracy.**

Most VGI in OpenStreetMap consists of vector data captured using consumer-grade GNSS receivers. Most users probably expect to use the data at scales greater than 1:25 000 (ie: within urban areas), so all the sub-dimensions of positional accuracy are relevant here, for both *absolute* and *relative accuracy*: *planimetric* and *vertical accuracy*, and *geometric fidelity*. Quality assurance is primarily done by other contributors and users.

- **Thematic accuracy.**

This is likely to be a problem, particularly where users might not understand the taxonomy properly, especially for points of interest (ie: *classification correctness*). Much attribute data are contributed as well, so *qualitative attribute accuracy* and *qualitative attribute correctness* are also important. Quality assurance is primarily done by other contributors and users. OSM also allows folksonomies (contributors can classify their data as they like), but it has a taxonomy of preferred classes.

- **Semantic accuracy.**

This is not an issue for most of the data capture (ie: for the VGI), as that does not involve integrating different data sets to produce new ones, but is an issue for the analysis, which could involve aggregating, interpolating and smoothing data.

- **Temporal accuracy.**

In areas where there are active contributors, OSM data can often be more up to date than official data. In other areas, OSM data are likely to be reasonably current, because OSM has been collecting VGI for less than a decade. All the sub-dimensions are relevant here.

- **Completeness.**

Coverage is uneven, with better coverage where there are more observers: developed countries vs developing countries, urban areas vs rural areas, etc. Ironically then, Figure 4.3 shows OSM data for Port Alfred, Eastern Cape, downloaded on 1 February 2012: the street network coverage then for Nkwenkwezi, an historically Black area, and for Station Hill, an historically Coloured area, are much better than that for the historically White areas such as East Bank, Kelley’s Beach and Forest Downs! It is difficult to determine where there are no data, except by comparison to other data (where available). OSM does arrange mapping parties to obtain data in unmapped areas, and/or to update or improve data.

- **Logical consistency.**

This is not addressed explicitly in the OSM documentation, but such quality problems would be detected by the OSM peer review processes.

- **Lineage.**  
The history of the data and some details of the observers are recorded. Trame & Keßler [2011] have used the OSM lineage records to produce ‘heat maps’ of edits, to learn about the edit, co-edit and tagging patterns in OSM.
- **Dependence on purpose & context.**  
Most contributions are probably to provide data for generic use at reasonably large scales. A record does not have metadata indicating the abilities of the contributor, though as it does identify the contributor, one could use all their records to assess their abilities in comparison to other contributors.
- **Non-involvement in standards.**  
As OSM has evolved, its standards have been developed with extensive involvement from its contributors and others.
- **Anonymous contributions.**  
OSM no longer allows anonymous contributions. Contributors are identified by a user name, but have to provide a valid email address to OSM.
- **VGI bias.**  
Potentially, a contributor could provide biased or false data to support their particular agenda. OSM’s primary defence against this is the sheer number of active contributors providing peer review, though one does need to consider the danger of proof by repeated assertion [Keeler 2011]. Du Plooy [2012] found that five contributors provided 73% of the data for both of his two test sites, Pretoria North and Bronkhorstspuit in Gauteng, which are nearly 60 kms apart. Mooney & Corcoran [2014] found that the top 10% of contributors in their assessment performed over 90% of all object creations and edits. Contributors also seem to be more active in their ‘home’ regions [Zielstra *et al* 2014].
- **Qualitative aspects.**  
OSM does allow folksonomies, but these are dealt with by providing a mapping to the preferred OSM taxonomy.

### 6.8.3 Tracks4Africa

*Tracks4Africa (T4A)* is a repository and a Web site initially of roads and tracks in Africa, but also of places of interest, that essentially began in 2001, see Figure 6.8. The network data are contributed in the form of GNSS tracks, voluntarily and on their own initiative by individuals directly to the Web site, and hence are a classic form of VGI. Tracks4Africa synthesises the contributed data to produce the road networks (currently over 710 000 km), effectively using multiple entry for the quality assurance: the synthesised data are only produced when they can corroborate the contributions of several different people [Tracks4Africa 2016]. Only the synthesised data are made available, not the raw contributed data.

Tracks4Africa also bundles in data to promote eco-tourism (eg: accommodation, fuel availability, and attractions — currently, over 138 000 points of interest and over 32 000

## 6. Quality

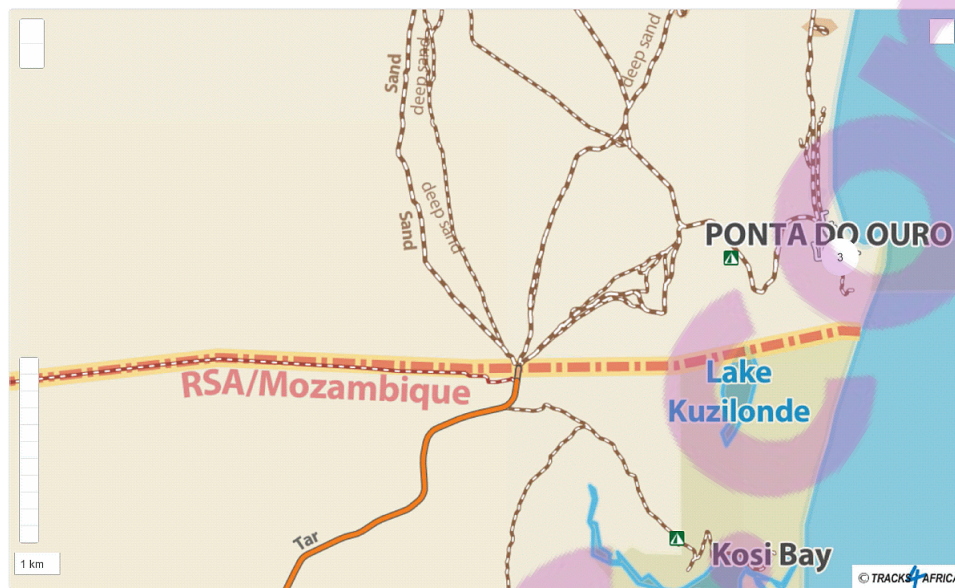


Figure 6.8: Tracks4Africa showing the Ponto Do Ouro area [Tracks4Africa 2016].

geocoded photographs), as obtained from its initiative, [Padkos 2016], and other sources. Tracks4Africa sells the data (updated twice a year), but the data are also available on Google Earth [Tracks4Africa 2016] and as part of the gROADS initiative<sup>16</sup> [gROADS 2014]. Tracks4Africa now also uses their data to make hardcopy maps, sold through retail outlets (eg: Tracks4Africa [2011]).

Nominally, Tracks4Africa is similar to OpenStreetMap, but it was selected for this analysis because it does have significant differences, such as a different business model (eg: selling data and publishing only synthesised road data) and its focus on Africa. Its headquarters are also in South Africa, in Stellenbosch. Marais [2010] studied the open innovation model used by Tracks4Africa and describes briefly its quality assurance process. However, while several scholars have used Tracks4Africa data, such as Roever *et al* [2013], currently there does not appear to have been anything published on the quality of the data in Tracks4Africa. For Tracks4Africa, please see also Section 8.3.2.2.

- **Positional accuracy.**

Most VGI in Tracks4Africa is vector data captured using consumer-grade GNSS receivers, but the data are only published after being confirmed by data from other contributors. Hence, even though most users probably expect to use the data at scales greater than 1:25 000 (ie: within urban areas), all the sub-dimensions of positional accuracy are relevant here, for both *absolute* and *relative accuracy*, though less of an issue unless there is systematic bias across the different data sets: *planimetric* and *vertical accuracy*, and *geometric fidelity*. Quality assurance is also done by other contributors and users.

<sup>16</sup>Global Roads Open Access Data Set, Version 1, developed under the auspices of the Global Roads Data Development Task Group of CODATA, the Committee on Data for Science and Technology of the International Committee on Science (ICSU) [gROADS 2014].

- **Thematic accuracy.**

While contributors can use their own taxonomy and attributes, they have to provide descriptions. The data are then classified by the T4A administrators before being published. So, *quantitative accuracy*, *qualitative accuracy* and *classification correctness* are all dependent on the diligence of the contributor and the quality of the descriptions they provide. Quality assurance is also done by other contributors and users.

- **Semantic accuracy.**

The only data that are published are aggregated. However, semantic accuracy is unlikely to be a significant problem, because the aggregation is done by the T4A administrators using data from independent contributors. *Aggregated shape fidelity* relates primarily to the shape of the roads while *aggregated type selection* relates primarily to how the points of interest are described by the contributors and allocated to the relevant classes by the T4A administrators.

- **Temporal accuracy.**

T4A data are likely to be more *current* than official data for many of the parts of Africa where T4A data have been captured, which are areas where over-landers, 4x4 enthusiasts and the like travel. In other areas, T4A data are likely to be reasonably *current*, because T4A has been collecting VGI for less than a decade. Because much of the data are captured using GNSS receivers, the *accuracy of a time measurement* and *temporal consistency* are likely to be high. The *temporal validity* will be fairly high, because Tracks4Africa only started in 2001, but the *updating efficiency* will be highly variable, depending on where contributors travel.

- **Completeness.**

Coverage is uneven (even in its published paper maps), with better coverage where its contributors live or travel. It is difficult to determine where there are *missing data of unexpected data*, except by comparison to other data (where available).

- **Logical consistency.**

This is not addressed explicitly in the T4A documentation, but such quality problems would be detected by the administrators and the T4A peer review processes, for all of *conceptual*, *domain*, *format* and *topological consistency*.

- **Lineage.**

The history of the data and some details of the observers are recorded, so all the *source information* and *process step information* should be available from the T4A administrators.

- **Dependence on purpose & context.**

Most contributions are probably to provide data for generic use at reasonably large scales.

- **Non-involvement in standards.**

T4A provides standards for field data collection, which are updated with the T4A community. The standards also promote safe, environmentally conscious and ethical cross-border travel.

## 6. Quality

---

- **Anonymous contributions.**

Contributors have to register with T4A, providing contact details.

- **VGI bias.**

Potentially, a contributor could provide biased or false points of interest to support their particular agenda, but this would be detected by the administrators and the T4A peer review processes. It is also possible that large parts of the T4A data set could come from very few contributors, or from several different contributors who were travelling in convoy.

- **Qualitative aspects.**

Allows folksonomies, but these are reclassified by the T4A administrators before publication.

### 6.8.4 Summary of the quality of these three repositories

I have assessed here three repositories of volunteered geographical information (2nd South African Bird Atlas Project, OpenStreetMap and Tracks4Africa) against the seven dimensions of quality (see Section 6.5) and in terms of five challenges for the quality of VGI (see Section 6.7.2).

It is clear that all three repositories have procedures in place to check the quality of the data, showing that the implementers of VGI repositories are aware of the importance of quality and the related challenges. The procedures of OSM are the most extensive and transparent, while SABAP2 depends on its vetting committees. In all three cases, uneven coverage is a challenge. Finally, this work confirms how difficult it is to assess data quality and that the quality assessment depends on the intended usage of the data.

While I assessed only three repositories here, they are quite different and hence I can offer some suggestions and recommendations regarding the usability of VGI in general.

- The problem of *classification correctness* might well be more important than realised for the usefulness of VGI, as even experienced contributors might be using classification systems that are out of date, or their own peculiar folksonomies. Hence, educating contributors on the correct classification system is very important.
- The *updating efficiency* and *completeness* can be very uneven, depending on the availability of contributors and the volumes they can contribute.
- The availability of suitable and detailed *metadata* remains a problem, with no obvious solution, though work is being done on the automated creation of spatial metadata [Kalantari *et al* 2010; Olfat *et al* 2012].
- It is very important to involve contributors in the *development of standards and protocols*, to encourage participation, reduce opposition to the standards or protocols and improve the data.
- The best defence against *biased* or *false data* would appear to be obtaining multiple records from independent contributors and/or peer review, but one must be aware

of the danger of *proof by repeated assertion* [Keeler 2011].

It will be useful to apply this analysis to other VGI repositories and to spatial data infrastructures (SDIs). It will also be useful to conduct a more in-depth study of the quality of the data in these three repositories (and others), especially for SABAP2 and Tracks4Africa (as many others have already assessed at OpenStreetMap), though perhaps with a specific application of the VGI in mind.

## 6.9 Using quality to classify geospatial data

### 6.9.1 Quality in taxonomies

Aspects of quality could be part of a taxonomy of geospatial data in general, or of VGI in particular. These could be the rigour of the screening based on quality; the availability and type of metadata; the extent to which liability for the data is accepted; and/or the quality dimensions and sub-dimensions described in Section 6.5. The quality of the VGI contribution could then be assessed based on its associated class in the taxonomy. If a taxonomy of VGI-based repositories has inadequacies, that suggests that there is a deficiency in the VGI quality itself — perhaps in terms of its completeness and/or in terms of not meeting needs of certain users.

Several attempts have been made to develop taxonomies of VGI, such as Coleman *et al* [2009]; Budhathoki *et al* [2009]; Castelein *et al* [2010], see Section 8.4. As mentioned above in Section 6.3, *quality* in general is one of the *issues* in the taxonomy of Budhathoki *et al* [2009] and is considered part of the *digital content policies* aspect of UGC by Wunsch-Vincent & Vickery [2007], but is not part of their taxonomy.

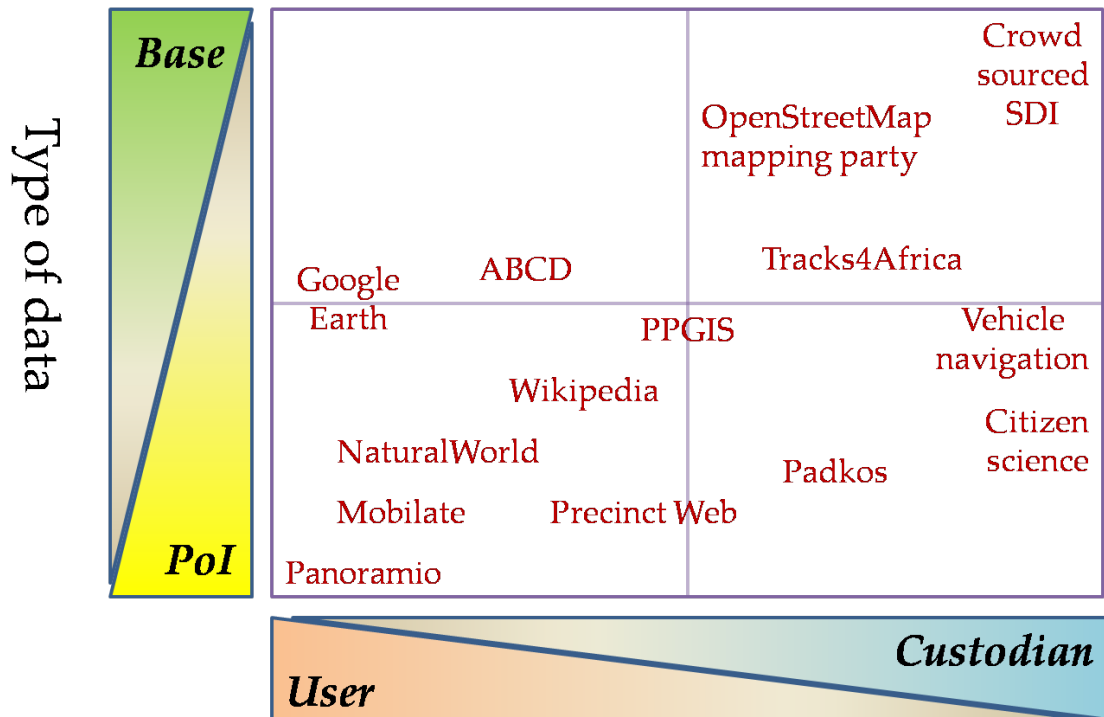
This could be taken further by using some or all of the dimensions or sub-dimensions of quality for a taxonomy of VGI. Effectively, I have done this above in Section 6.8 for three VGI repositories. However, to be practical, the feature types (or classes) in the taxonomy would need to be rendered in a suitable form for classification, as detailed in Section 2.4. The key problem with doing this, though, is ensuring that each feature type is unique. To achieve this, the feature types would need to be such that they can be ranked from worst to best. The South African Statistical Quality Assessment Framework (SASQAF) [Statistics South Africa 2008] does this, combining quality dimensions to provide four levels of certification, see Section 6.6.4.

### 6.9.2 Responsibility for specifications *vs* types of data

Figure 6.9 presents two possible dimensions of a taxonomy for VGI that are particularly important for understanding quality issues regarding VGI [Cooper *et al* 2011a]. On the horizontal axis, we have the continuum of responsibility for determining the specifications for the data, ranging from a user on the left (effectively, near-anarchy) through to an official data custodian on the right (with tightly controlled specifications). The vertical axis ranges from base data at the top to points of interest (PoIs) at the bottom — it

## 6. Quality

is a continuum because classifying data as base or PoIs can depend on one's perspective and applications of the data. For a discussion on the nature of base data and PoIs, see Section 2.3.



### Determination of the specifications

Figure 6.9: Types of VGI from the perspective of quality [Cooper *et al* 2011a].

The grid in Figure 6.9 is populated with examples of repositories of VGI, that are described in Section 8.3 and used in Chapters 8 and 9 for assessing various taxonomies. In the bottom left of the figure is *Panoramio* [Panoramio 2016], which has arbitrary photographs of places added to Google Earth [Google 2016a], sometimes incorrectly labelled and positioned. *Google Earth* itself is an undifferentiated repository of data, spanning both base data and PoIs, and including both VGI and data from official sources.

Top right is a *crowd-sourced SDI* (which probably does not yet exist), where users contribute data according to a tight specification from the custodian, who would then subject the VGI to their usual quality assurance processes. Also in the top-right quadrant are repositories of VGI that are primarily base data, particularly road and street networks, that are subject to fairly tight specifications (eg: *OpenStreetMap* and its *mapping parties* [OpenStreetMap 2016]) and/or rigorous quality assurance (eg: *Tracks4Africa*, which uses statistics to produce a best fit from multiple contributions for each road, street or track segment [Tracks4Africa 2016]).

On the lower right are repositories with tightly-defined specifications for PoIs, such as *in-vehicle navigation systems* (providing real-time traffic densities) and *citizen-science* projects

such as SABAP2. In the lower-left quadrant are the likes of *Mobilitate*<sup>17</sup> [Mobilitate 2015], for logging complaints about service delivery in South Africa, *NaturalWorld* [NaturalWorld 2016] and *Wikipedia* [Wikimedia 2016], which relies on open peer review of its articles. *Precinct Web* [Precinct Web 2016] has more rigorous specifications for mapping crime in South Africa and *Padkos* [Padkos 2016] is Tracks4Africa's site for PoIs (accommodation, restaurants, shops, etc).

The specifications and data for *asset-based community development (ABCD)* will evolve as the community discovers what is important to them. The *public participatory geographical information system (PPGIS)* is in the middle because it includes contributions from both custodians (eg: local authorities) and community members (the VGI), and both base data and PoIs. Unsurprisingly, the top-left part of the grid is largely empty, because base data are widely used and hence need specifications.

This analysis shows the diversity of VGI and that it is not possible to adopt a rigid and narrow perspective on VGI. Clearly, some VGI is definitely not suitable for an SDI, possibly even with extensive post-processing and quality assurance (eg: the photographs on *Panoramio*). However, there is already much VGI of a quality comparable to official or commercial data (eg: see Haklay [2010]), as discussed above in Section 6.7.

## 6.10 Standards for the quality of geospatial data

ISO 9000 [2005] and its related standards (eg: ISO 9001 [2008] and ISO 9004 [2009]) are for an organisation's quality management system and do not specify requirements for goods and services, such as geospatial data. Rather, they deal with the seven quality management principles of customer focus, leadership, engagement of people, process approach, improvement, evidence-based decision making and relationship management.

ISO/TC 211, *Geographic information/Geomatics*, has developed several standards for the quality of geospatial resources (particularly data and services) that can be used within the ISO 9000 framework. The following has been summarized from the ISO/TC211 Standards Guide [Roswell 2009].

- ISO 19113:2002, *Geographic information — Quality principles*.  
This standard provides principles for describing the quality of geospatial data, concepts for handling such quality information, and an approach for organising such information. However, it does not attempt to define a minimum acceptable level of quality for geospatial data. ISO 19113 uses data quality elements and sub-elements to describe the quality of data. The following are the elements: *completeness*, *logical consistency*, *positional accuracy*, *temporal accuracy* and *thematic accuracy*. As can be seen, these match most of the commonly used dimensions, described in Section 6.5. ISO 19113 also allows additional data quality elements to be created to describe other aspects of quality.

<sup>17</sup>Please note that while Mobilitate no longer exists, it was still valid to use it for the analysis done in this thesis, as it represents a type of Web service and geospatial data repository. A similar service started recently in South Africa is LocalBlock [2016].

## 6. Quality

---

- ISO 19114:2003, *Geographic information — Quality evaluation procedures*.  
This standard provides a framework of procedures for determining and evaluating the quality of geospatial data, using the data quality principles defined in ISO 19113 [2002], and for evaluating and reporting data quality results. The quality evaluation can be direct (comparison with reference data) or indirect (inferred or estimated from the lineage or other sources).
- ISO/TS 19138:2006, *Geographic information — Data quality measures*.  
For each data quality sub-element defined in ISO 19113 [2002], this specification defines a set of multiple measures of data quality, from which one can select depending on the type of data and their intended purpose.
- ISO 19157:2013, *Geographic information — Data quality*.  
This standard combines, updates and replaces ISO 19113 [2002], ISO 19114 [2003] and ISO 19138 [2006]. ISO 19157 [2013] caters for most of the commonly-used quality dimensions, as discussed in Sections 6.5 and 6.6.

The Open Geospatial Consortium Inc (OGC) does not have a standard or specification for data quality itself, because it uses the above standards from ISO/TC 211. Previously, OGC considered having spatial data quality certification as part of its interoperability programme.

Table 6.2 shows how the quality dimensions can be used with VGI. Camboim *et al* [2015] used two of the dimensions to assess VGI in Brazil, looking at *completeness* and *updating efficiency*, and Dorn *et al* [2015] used *completeness* and *classification correctness* in Germany.

Surprisingly, Goodchild [2008b] claimed that these standards insisted “that data quality is an attribute of a single data set”. However, he did point out that one needs “to know the relative data quality of pairs of data sets that are being integrated through mashups and other means”. He uses the term *binary metadata* for describing “the ability of two data sets to work together, since such information cannot be deduced from the unary metadata records” [Goodchild 2008b].

### 6.11 Summary and looking ahead

While this chapter provides the setting for subsequent chapters, it also makes important contributions as part of my research and this thesis. Drawing on Chapters 2, 4 and 5, this chapter has provided details of the different aspects of the quality of resources, the four stages for recognising the quality of a resource, GNSS errors, the dimensions of quality, challenges for the quality of VGI, the quality of three VGI repositories, quality and classification, and standards for the quality of geospatial data.

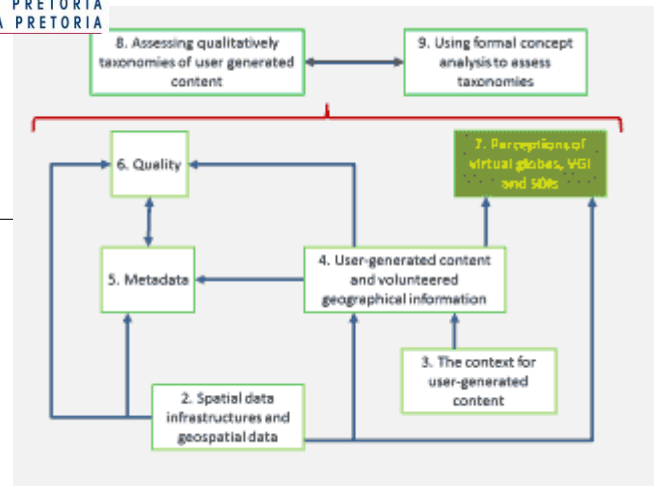
The major original contribution that I have made that is presented in this chapter is in identifying the **four stages for the recognition of the quality** of a resource in general. The key contributions that I have made here are:

- Presented the dimensions and sub-dimensions of quality cohesively;

- Identifying some challenges for VGI;
- Assessing three VGI repositories against the quality dimensions and quality challenges; and
- For several VGI repositories, mapping the responsibility for their specifications against the types of data they contain.

A common objection raised against VGI is uncertainty over the quality of the VGI. This is now seen clearly in Chapter 7, which has been published in a special issue of a journal [Cooper *et al* 2010a]. It reports on the results of a survey conducted through a questionnaire (included in Appendix A), of geographical information professionals in Africa in general and in South Africa in particular, concerning their perceptions of virtual globes, volunteered geographical information and spatial data infrastructures.

\*\*\*\*



## Chapter 7

# Perceptions of virtual globes, volunteered geographical information and spatial data infrastructures

### 7.1 Overview of the chapter

The contents of this chapter were published as a special issue of the Canadian journal, *Geomatica* (the journal of geospatial information science, technology and practice), on *Volunteered Geographic Information (VGI)*. With the exception of the first section of that paper [Cooper *et al* 2010a] (which is presented in far more detail in Chapters 1, 2 and 4 of this thesis), this chapter is more detailed than that paper and presents the results of the two surveys separately. The questionnaire used is included in Appendix A.

While the concept of a *spatial data infrastructure* has been around for a long time (since the beginning of the 1960s, with the *Canadian Geographic Information System (CGIS)* [Tomlinson 1988]) and while the labels *spatial data infrastructure* or *SDI* have been used since 1990 [National Academy of Sciences 1990], that does not mean that the concept is well understood — particularly in Africa, which does not have a good track record of building and sustaining SDIs [Makanga & Smit 2008]. The concepts of *virtual globes* and *volunteered geographical information* are much more recent, as is discussed above in Sections 4.5 and 2.10. Access to virtual globes through the Internet is constrained by the speed, reliability and

---

## 7. Perceptions of virtual globes, VGI and SDIs

---

cost of bandwidth, which are problems across Africa [6DISS 2005; Zennaro *et al* 2006]. On the other hand, given the limited availability of up-to-date official geospatial data for much of Africa, one might expect VGI to be common in Africa, as communities map their areas (eg: the first map of Kibera in Nairobi, Kenya, is crowd-sourced VGI [Mulupi 2011], see Section 8.4.2) or as companies exploit the gap (eg: Tracks4Africa [2016]).

Hence, it is useful to assess what the perceptions actually are of these concepts. The major original contribution that I have made that is presented in this chapter is:

- To geographical information science, conducting a survey through a questionnaire of geographical information professionals concerning their perceptions of virtual globes, volunteered geographical information and spatial data infrastructures.

Responses to the questionnaire were obtained from professionals in Gauteng, South Africa (see Section 7.4), and in Africa in general (see Section 7.3). These perceptions are important because they determine the future use of VGI and virtual globes in these communities. This chapter includes my analysis of the results from the survey, noting the similarities and differences, and describing the issues arising that warrant further investigation: see Section 7.5. It would also be useful to apply the survey elsewhere, as discussed in Section 7.6.

### 7.2 Background to the questionnaire

During April 2009, I compiled a questionnaire in English on the use of volunteered geographical information in a spatial data infrastructure, with some inputs from colleagues. A copy of the questionnaire is included in Appendix A. The two-page questionnaire was printed on a single A4 sheet, double-sided. This limited the number of questions that could be asked and was intended to ensure that an individual's responses could not be separated. Unfortunately, it was not made obvious on the first page that there were questions on the reverse, and several respondents did not answer any of the questions on the reverse.

With permission from the United Nations Economic Commission for Africa (UN ECA), a copy of the questionnaire was circulated at the end of April 2009 at the meeting of the Geoinformation Subcommittee of UN ECA's Committee on Development Information, Science and Technology (CODIST), held in Addis Ababa, Ethiopia. The membership of CODIST-Geo consists of senior representatives of relevant government departments, such as national mapping agencies, topographical surveying departments and cadastral surveying departments. However, the meetings also include observers from academia, non-government organisations, the private sector, international organisations, and from outside Africa. Some of these observers are also designated as UN ECA resource persons, because they make presentations or act as rapporteurs.

The reason for selecting the CODIST meeting was that I had been invited to make a presentation at the CODIST Plenary (which included delegates from all three of CODIST's sub-committees) as a Discussant on behalf of CODIST-Geo. This questionnaire then drew

## 7. Perceptions of virtual globes, VGI and SDIs

---

on some of the ideas in the paper I presented, entitled *Geoinformation perspectives on innovation and economic growth* [Cooper 2009b]. It was also an opportunity to gauge opinions from other African countries.

Unfortunately, while about 100 paper questionnaires were circulated to CODIST-Geo (and an electronic version given to selected delegates on request), only 14 were completed and returned to me (13 at the meeting and one emailed later to me). This was not entirely unexpected, however, as by its nature, the questionnaire had an advocacy component because VGI is a new concept and possibly unknown to some of the delegates. Some delegates might have retained the questionnaires as a reference document for when they had returned home. Further, many national mapping agencies in Africa are constrained by lack of equipment, skills and funding — some are still restricted to manual cartography only. Hence, for them, virtual globes and VGI can represent threats to their sustainability and they might have been reluctant to respond to the questionnaire. Further, it was not possible to translate the questionnaire into French given the tight deadlines and this lack of a French version would have reduced the number of responses, as many of the delegates at CODIST-I were from Francophone Africa and some of them are not fluent in English.

I also circulated about 25 questionnaires at a meeting of the Gauteng Branch of the Geoinformation Society of South Africa (GISSA), hosted by the Ekurhuleni Metropolitan Municipality at their offices in Kempton Park on Friday, 19 June 2009. This meeting was selected because it was the first relevant local meeting I attended after CODIST-I and because I reprised my CODIST-I presentation for the local audience. Seventeen questionnaires were completed and collected at the GISSA meeting.

Given the limited number of responses received, it is not possible to draw any statistically-valid conclusions from the questionnaire. However, it was never the intention that these admissions of this questionnaire should provide empirical data. Rather, the purpose was to perform some qualitative research to gauge the opinions of informed persons interested in responding to the questionnaire. These responses could be used to refine the questionnaire so that it could be used to gather empirical data from which statistically valid conclusions could be drawn about some population, though that would probably be an expensive exercise.

In drafting the questionnaire, both free-text and multiple-choice questions were included deliberately, to see what effect they would have on the responses received. Free-text questions were used for questions 3 to 10 to minimize the bias of the questionnaire, especially as the disadvantages of virtual globes, geobrowsers, VGI and the lack of metadata might not be well known, and some of the respondents might not have considered their impact on official mapping. I believe that the responses have supported this.

In general, it appears that the responses to the free-text questions could be used to draft meaningful categories to convert these questions into multiple choice questions, but this would undoubtedly bias the responses.

The following definitions for a virtual globe and geobrowser were provided in the questionnaire:

## 7. Perceptions of virtual globes, VGI and SDIs

- A *virtual globe* provides masses of digital geographical information over the Internet, typically in the form of a globe.
- A *geobrowser* is the interface to a virtual globe, typically allowing users to zoom into the data, switch data layers on and off, create three dimensional views and add their own data (user generated content), such as geographical features (e.g. roads and places of interest), tags (with text or links to Websites) and photographs.

Perhaps the best-known example of a virtual globe is Google Earth.

While these definitions distinguish between a ‘virtual globe’ and a ‘geobrowser’, we pointed out in the introduction why the terms are sometimes used interchangeably to refer to Google Earth. Also, in the questionnaire, the two terms were treated as a single entity, eg: *What do you think of the quality of the data in virtual globes/geobrowsers?* Since Google Earth is by far the most widely used virtual globe, evident from the responses, it is most likely that they had Google Earth and its functionality for user generated content in mind when answering the questions on virtual globes/geobrowsers in the questionnaire.

### 7.3 Summary of the results from CODIST-I

The questionnaire was circulated to CODIST-Geo in the morning of Tuesday, 28 April 2009, and the completed questionnaires were collected during the week, though mainly on the Tuesday. The emailed response was received within a fortnight after CODIST-I — it was a much more detailed response and from a respondent who has clearly given the issues much thought. My plenary presentation was made late on the Wednesday afternoon, so it probably had no influence on the responses received. Some of the issues were discussed with some respondents before they completed the questionnaires, but these specific individuals are sufficiently well informed about virtual globes, geobrowsers and VGI that their responses were not influenced significantly by such discussions.

One weakness in the questionnaire highlighted by these responses was that the following question was misinterpreted:

*7. What do you think of the documentation of the data (ie: the metadata) in virtual globes/geobrowsers?*

The intention of this question was to assess what the respondents thought of the quality of the metadata currently available in virtual globes/geobrowsers, but some interpreted this question as asking if they thought that metadata was necessary *per se*.

The only respondent to submit an anonymous response was the one who completed only the first page of the questionnaire (up to question 11. *Do you think that the legislative and policy environment in your country encourages or stifles innovation in the field of geographical information?*), and hence missed the request for their details.

The following is an assessment of the responses received. Questions 3 to 10 inclusive required free-text responses and to preserve the privacy of the respondents, these answers

## 7. Perceptions of virtual globes, VGI and SDIs

---

have been mixed up and interpreted — hopefully correctly! Questions 11 to 22 were multiple-choice questions.

### 1. *Country of current residence*

Responses were received from 10 African countries, one European country and one Asian country. Most of the African responses were from southern Africa, but there were also responses from west, north and east Africa. While this does not make the responses representative of Africa, of course, they do at least provide a bit of regional variety.

### 2. *Economic sector in which employed*

Most of the respondents are in government, but three are in academia (one also in the private sector) and one is at a non-governmental organisation (NGO).

### 3. *Main advantages of virtual globes and geobrowsers*

Several respondents identified the main advantages as being quick and easy access to free data, the ability to share data (particularly of current global events), and the low skills required to access the data. Some also identified the role of virtual globes/geobrowsers in assisting visual planning and quick decision making, particularly in allowing broader participation from earlier on, facilitating multiple views of the situation and promoting feedback and dialogue — replacing the moralistic rhetoric of ‘ought’ with a technical analysis of ‘is’.

Some felt that the data were valuable and ranged from “relatively good” to “more precise and tested”, which would seem to contradict some of the responses to the next question. A virtual globe also provides a unique global reference and promotes the democratization of data by allowing technical analyses countervailing those of intelligence and other government agencies to shift the epistemic balance of power between civil society and the state — for example, by using VGI and the satellite imagery on virtual globes as resistance to military secrecy. Finally, virtual globes/geobrowsers have brought geographical information to lay people, allowing them to play with the data for fun, such as engaging in virtual tourism, searching for interesting things<sup>1</sup> or make subversive mash-ups.

### 4. *Main disadvantages of virtual globes and geobrowsers*

The key disadvantage identified is the uncertainty over the legitimacy and quality of the data (can one trust what is on the virtual globe?), because of the lack of moderation over what is added and the lack of metadata which could be used to determine the quality of the data (eg: currency or positional accuracy). It also means that one cannot identify when data have been removed or edited at the behest of a government or someone else, in an effort to delude the public. The perception held by the organisation owning the virtual globe of what is important in terms of the currency and resolution of the data might be tailored to their perception of market potential, which might not gel with the public interest<sup>2</sup>. Virtual

<sup>1</sup>For example, the scale model built near Huangyangtan, China, of a disputed border area in Tibet [Haines 2006].

<sup>2</sup>For example, the one might perceive currency as being more important than resolution, while the other might favour resolution over currency.

## 7. Perceptions of virtual globes, VGI and SDIs

---

globes/geobrowsers allow the visualizations of lay people to enter the public discourse and affect decision-making, which raise the difficult value questions of “who has a legitimate voice?” and “whose visualisation is right, or more legitimate?”

The respondents also raised the issues of invasion of privacy (making surveillance available to everybody), the exposure of sensitive sites (this is not just in terms of national security, of course, but also applies to cultural and environmental sites<sup>3</sup>), the risk of the data being used by vandals or criminals, and the security of the data placed on the virtual globes. Geobrowsers have limited functionality. One respondent considered a disadvantage to be the availability of free data, presumably because of the threat it poses to national mapping agencies.

Finally, of course, to be able to use a virtual globe/geobrowser, one needs electricity, a computer and connectivity — never mind reasonable bandwidth — and these are luxuries in many African countries.

### 5. *Main advantages of user generated content in a virtual globe/geobrowser*

The only common answer was that it allows ordinary people to contribute data quickly and easily that are then globally available. Other issues that the respondents mentioned were that VGI can give a geographical context to imagery; it encourages ordinary people to become interested and add their local knowledge to suit their needs; it adds value to the generic viewer which can benefit other users; it reflects an individual’s ideas in the information exchange; it’s an unlimited source of data; facilitates quick generation of user-defined answers and easy customization; the data are unedited (*did the respondent mean uncensored?*); and includes three-dimensional data.

### 6. *Main disadvantages of user generated content in a virtual globe/geobrowser*

Most of the respondents cited concern over the veracity of the data and hence over knowing which data sets to use — unmoderated, unverified, uncontrolled, subjective, inadequate precision, the lack of a common standard, and that data might be misunderstood by others. Some also expressed concern over the longevity of the data (they are disposable), that VGI might “pollute” (ie: obscure or replace incorrectly) the base data, and that the ability to propagate VGI is open to abuse. The limited availability of fast connectivity denies many the opportunity to contribute VGI (which biases the available VGI). Different respondents said that it grows a user community beyond the traditional GIS community and provides quick access and definition of user-defined uses — presumably, the common problem here is that the user-generated content could be produced carelessly and without understanding of key issues, such as geo-referencing. Attention shifts from what happens inside a single organisation, to what happens in the new social system of geo-information production. As a result, the right to define and judge the value of the geo-information being co-produced is distributed among all co-producers; and new rules and standards are required to take into account the values of the equity of volunteers, security, community building and privacy, in the evaluation of the performance of the new production system.

---

<sup>3</sup>Ruthless collectors exploit the data to steal fossils and cultural artefacts, for example.

## 7. Perceptions of virtual globes, VGI and SDIs

---

### 7. Documentation of the data (metadata) in virtual globes/geobrowsers

As mentioned above, some respondents misinterpreted this question as asking if they thought that metadata was necessary, and they obviously did. Generally, the other respondents felt that the available metadata was the biggest shortcoming of virtual globes/geobrowsers, was inadequate, incomplete, obsolete, not complying with international standards, contained errors (spelling and misidentification) and/or with the currency and resolution of images reflecting perceptions of market potential not of public interest. However, some respondents felt that the metadata was OK for most practical purposes.

### 8. Quality of the data in virtual globes/geobrowsers

The responses varied from “very poor” through “acceptable” to “high”, though depending on the application! Perhaps this reflects that the data (both the base imagery and the VGI) come from disparate sources with variable degrees of quality, with the imagery being considered to be better. Specific issues raised are that the data for the American continent are better than for the African continent, the age of the data sets is ambiguous, and information on the quality of the data is not available. The data need to be peer reviewed for them to be used for scientific purposes, but not necessarily for obtaining opinions.

### 9. Current impacts of virtual globes/geobrowsers on official mapping

This is a topic of great research interest, because of the issues raised by the other questions. At the moment, the impact on official mapping is considered to be low to none. The main impact was identified as being in the early stages of the mapping cycle — planning, viewing places, as a backdrop for vector data and preparing working documents. Virtual globes/geobrowsers are attractive for both experienced users and novices, and hence could reduce the importance of official mapping, but could also help official mapping as their use and understanding improves. There are also some issues of privacy with the data being opened to the public.

### 10. Impacts of virtual globes/geobrowsers on official mapping through to end of 2014

Again, a topic of great research interest. Most of the respondents felt that the impact would be positive: to be used more than now for research, planning and perhaps updating other maps; to help gain access to new data; to help disseminate new products; to promote geo-information; to provide good access to geo-information; and to facilitate instant decision making by top level officials in government. They could also supplement the national mapping series — or they could reduce the importance of official mapping. One respondent felt they would have very little impact on national mapping agencies, but could assist thematic mapping. Some felt virtual globes/geobrowsers would impact on defining the mapping strategy and in planning and execution of mapping projects, but without saying if this would be negative or positive. To have an impact, the information would have to be updated.

### 11. Do you think that the legislative and policy environment in your country encourages or stifles innovation in the field of geographical information?

All respondents had an opinion, with four feeling that legislation and policy encourage innovation with geographical information, three feeling that they stifle innovation (one added that more effort is needed) and seven feeling that they neither

## 7. Perceptions of virtual globes, VGI and SDIs

encourage nor stifle innovation.

12. *Do you think that the legislative and policy environment in your country encourages or stifles the development of spatial data infrastructures (SDIs)?*

A positive response, with eleven saying that legislation and policy encourage SDIs and two saying neither encourage nor stifle. This is not surprising as South Africa was one of the first countries in the world to have an SDI Act [South Africa 2003], and several other African countries have been following suit.

13. *Do you think that the legislative and policy environment in your country encourages or stifles the development of, use of, and adherence to, standards?*

A fairly positive response, with seven replying that legislation and policy encourage standards, four replying for neither, one replying for stifles and one did not know. However, Africa has a very limited participation in international standards generating bodies. South Africa has been the only active African participant in ISO/TC 211, *Geographic information/Geomatics*, though Morocco has sent a delegate to one Plenary. Perhaps the legislation and policy need to be backed up with financial support?

14. *How well do you think the legislative and policy environment in your country deals with issues such as virtual globes, volunteered geographical information and open access to geographical information?*

Unsurprisingly, this resulted in a negative response, with two selecting very well, two selecting neither (one respondent marked both very well and adequately), but six selecting poorly and four selecting not at all.

15. *Access to a virtual globe/geobrowser at home*

Eight respondents have access to a virtual globe/geobrowser at home and five do not.

16. *Access to a virtual globe/geobrowser at work*

Ten respondents have access to a virtual globe/geobrowser at work and three do not. Overall, all those who have access at home also have access at work. Three do not have access at either home or work — this might surprise people from outside of Africa as the respondents are from the wealthier “classes” in Africa, but it does not surprise me. A key problem is the extraordinarily high costs of Internet access across Africa, because of all the telecommunications monopolies, which results in access costing over one thousand times what it costs in Europe, North America and North-East Asia [6DISS 2005; Zennaro *et al* 2006]. Hence, even if these respondents have access to the Internet, a resource such as a virtual globe consumes too much bandwidth and is either prohibitively expensive to use — or is even impossible to use because it is so slow and one is likely to lose the connection before one gets any results.

17. *Use of a virtual globe/geobrowser for personal purposes*

Seven of the respondents use a virtual globe/geobrowser for personal purposes, and six do not.

18. *Use of a virtual globe/geobrowser for work purposes*

## 7. Perceptions of virtual globes, VGI and SDIs

Eight of the respondents use a virtual globe/geobrowser for work purposes, one does sometimes, and four do not (including one with access at work who does not use it at all). One respondent said they did not have access to a virtual globe/geobrowser from home or work, yet they still used one for work purposes. Clearly, they would then use a virtual globe/geobrowser at a friend's house, at an Internet café, at a conference such as CODIST-I<sup>4</sup>, or the like. This is a clear indication of the limited availability of the Internet and Internet-based services in Africa, because these respondents are senior government officials or the like.

### 19. *The virtual globes/geobrowsers used*

Eleven of the respondents use Google Earth. One respondent also uses both NASA World Wind and Open Street Map. This is an indication of the dominance of Google Earth, both in actual use and in perceptions of what a virtual globe/geobrowser is.

### 20. *The main reasons for using a virtual globe/geobrowser*

The respondents could select several options if they so chose, with the results being as follows in Table 7.1:

Table 7.1: Reasons for using a virtual globe/geobrowser (CODIST respondents)

Reasons	Responses	Comments
Travel planning (work or leisure)	4	I expected this to be more popular!
Providing a geographical context to news items	1	The low response might be an indication of limited bandwidth, in that a user would not use a news Web site and a virtual globe simultaneously.
Accessing data for work purposes	6	This option was possibly badly worded as it was meant to see who used a virtual globe or geobrowser for specific project work, rather than used them for work purposes in general (eg: travel planning).
General curiosity	6	Unsurprisingly, a common activity.
Publishing your data	1	Given the responses to other questions, this correlates well with the low active use of virtual globes.

*Continued on next page*

<sup>4</sup>However, the bandwidth at CODIST-I was surprisingly limited, far worse than it had been at previous CODI meetings, so it is unlikely that any delegate used a geobrowser there for anything significant.

## 7. Perceptions of virtual globes, VGI and SDIs

Reasons	Responses	Comments
Reconnaissance for work purposes	6	This question was meant to gauge the use of the data on virtual globes for planning work activities, so the response is surprisingly high, given the other responses above. This option might have been confused with <i>travel planning</i> , which would then be an example of the very common weakness of brevity in questionnaires!
Providing a geographical context to correspondence from friends and family	0	The low response rate correlates well with that for <i>providing a geographical context to news items</i> .
Backdrop for other geographical data	1	This low response rate correlates well with that for <i>publishing your data</i> , because of the cost of maintaining the Internet link to the virtual globe to use it as a backdrop.
Armchair travelling	0	Surprisingly, no one selected this option, but it does overlap with <i>general curiosity</i> .
Searching for data	4	Clearly, this option could be considered to overlap with all the others, but it is likely that the respondents interpreted it to mean searching for data that they could download.
Other (please specify)	0	No responses, but only a heavy user is likely to give a response here.

Of course, there are some overlaps between these categories, such as between *general curiosity* and *armchair travelling*. This was deliberate, to ensure that the questionnaire covered what I anticipated would be the common uses of virtual globes and geobrowsers.

### 21. Use of VGI in a virtual globe/geobrowser

Of those who use a virtual globe/geobrowser, four use VGI and six do not. Of course, a key issue with exploring this issue is how easy it is for the casual user to identify VGI.

### 22. Use of a markup language in a geobrowser

Unsurprisingly, as it would only be used by those contributing structured data, only two respondents use a markup language in a geobrowser, while nine do not.

## 7. Perceptions of virtual globes, VGI and SDIs

---

### 7.4 Summary of the results from GISSA Gauteng

The questionnaire was circulated during the morning (19 June 2009) and while my presentation was the last of the day (mid-afternoon), some of the respondents only completed the questionnaire during this presentation. Further, there was discussion of issues such as the quality of the data in a virtual globe during some of the other presentations. One of the other presentations was about using KML and Google Earth to deliver government data [Silberbauer & Geldenhuys 2008]. Hence, these discussions probably influenced some of the responses. Nevertheless, circulating the questionnaires at this GISSA meeting was a useful exercise, complementing the responses from the CODIST meeting, as most of the respondents were from the private sector and some of the respondents are active users of virtual globes and geobrowsers.

Unlike the CODIST responses, ten of the GISSA responses were anonymous, with five being because these respondents did not complete the second page, missing the request for their details. Only one respondent misinterpreted the following question:

*7. What do you think of the documentation of the data (ie: the metadata) in virtual globes/geobrowsers?*

The following is an assessment of the responses received.

1. *Country of current residence*

Unsurprisingly, sixteen of the respondents reside in South Africa, with the seventeenth declining to answer.

2. *Economic sector in which employed*

Eleven of the respondents are from the private sector, four from government (some from local government, unlike the delegates at CODIST-I), one from academia and one declined to say.

3. *Main advantages of virtual globes and geobrowsers*

As with the CODIST respondents, several identified the main advantages as being quick and easy access for people in the street to a wide range of free data that are relatively up to date (particularly imagery) — spatially enabling society and making the public spatially aware (a map is worth a thousand words). They also create awareness about GIS-related technology and make the technology available to the public and easy to use, providing an interactive exposure to geography. This then results in the public demanding better quality spatial data and reduces the commercial sales cycle for the technology and data. Knowing “where” is now just the beginning. However, they do require connectivity to be accessible.

Virtual globes and geobrowsers multiply the spatial-enablement efforts of others. They allow one to concentrate on the data one is trying to present, while leaving the fancy image serving and draping to the geobrowser. They allow engineers to do high-level planning, such as identifying possible corridors. Finally, of course, virtual globes and geobrowsers are fun!

4. *Main disadvantages of virtual globes and geobrowsers*

## 7. Perceptions of virtual globes, VGI and SDIs

Again, the respondents identified the main disadvantages as being uncertainty over the quality, accuracy, currency, consistency and reliability of the spatial data, because of the lack of metadata, as well as the need for Internet connectivity with high bandwidth. Another main disadvantage is that naïve users can place too much faith in the reliability and accuracy of the data, often using them as an “exact science”. Such users could feel that all they need is the virtual globe and geobrowser because they are so easy to use, posing a potential threat to commercial GIS software. However, several respondents identified the limited functionality of geobrowsers as being a disadvantage, such as with exporting data, using them with other systems (possibly proprietary), the lack of graphical tools (eg: snapping to existing geometry), and requiring the purchase of the commercial version of the geobrowser to be able to upload data. There is uncertainty over whether higher resolution data are better than up-to-date data.

Some corporate computer centres don’t like installing the software (presumably because of bandwidth issues and corporate policies). One respondent felt that the virtual globes and geobrowsers were increasing the gap between the computer literate and computer illiterate. Another respondent did not know of any disadvantages.

### 5. *Main advantages of user generated content in a virtual globe/geobrowser*

The responses were similar to those from CODIST, with the most common advantage being that everyone is now able to contribute and share their spatial data (and maps and knowledge) and add value to other data sets. This adds local knowledge and enables the “wisdom of the crowd” to make its way into applications. Presenting results with an imagery backdrop (the cosmetics) is a “wow” factor. Several also mentioned the advantage of data being made available freely and anyone then being able to participate in a global community by looking for spatial data by foraging for them in a visual landscape.

Other issues mentioned are that VGI can be suitable for the needs of many users and fit for their purposes; one can see and experience areas of interest from one’s desk; the increasing volume of data becoming available; the VGI facilitates verification; the VGI allows a free, easy-to-use application to act as a GIS; and the large “help desk” effectively available through the community using the virtual globes and geobrowsers.

### 6. *Main disadvantages of user generated content in a virtual globe/geobrowser*

Unsurprisingly, considering the discussions during the presentations at the meeting, half the respondents gave the main disadvantage as being the quality (accuracy, currency, trustworthiness) of the VGI and the uncertainty of the quality (how does one verify the data?). Then, those users not aware of the quality issues could have the attitude “I saw it on the Internet so it must be true”.

Other issues mentioned are limitations on uploading data; the VGI might not meet one’s perspective; the required data might not be available; the VGI might be dependent on transient details in the background imagery and might lose its context when the imagery is updated<sup>5</sup>; security; and the lack of support for applications.

<sup>5</sup>The classic problem of the incremental updating and versioning of base spatial data sets [Peled & Cooper

## 7. Perceptions of virtual globes, VGI and SDIs

Again, one respondent could not think of any disadvantages.

### 7. Documentation of the data (metadata) in virtual globes/geobrowsers

Most of the respondents were unimpressed with the quality, quantity, depth, currency and verification of the metadata and felt it should be improved and adhere to standards. One felt it was getting better. However, one respondent felt that the metadata was not relevant and two considered it to be generally very good and up to date. Two had not investigated the metadata.

### 8. Quality of the data in virtual globes/geobrowsers

The responses varies from “very coarse” (especially for road data) or “questionable” (especially positional accuracy), to “very good” or “high standard”, with most respondents rating the quality as being “fair”/“adequate” or better. Several respondents highlighted the need for the data to be maintained and updated regularly. Several also pointed out that they cannot assess the quality without there being adequate metadata and others pointed out that the quality required depended on the use and the scale, and how much one was prepared to pay for quality data (VGI tends to be free). One considered most of the data to be vague and not important for general users. The ownership of the data is also a problem.

### 9. Current impacts of virtual globes/geobrowsers on official mapping

Several respondents felt that virtual globes/geobrowsers were already changing official mapping for the better, such as by forcing them to be more consumer oriented; educating them to understand the value of information; and creating a greater awareness amongst the public of spatial data. One respondent felt that official mapping should be provided through a geobrowser. Another respondent acknowledged that the presentation at the meeting by Mike Silberbauer (see [Silberbauer & Geldenhuys 2008]) made them realise that virtual globes and geobrowsers have already had a significant impact! Several respondents pointed out that to have a real impact, the data need to be up to date and accurate — Maps4Africa<sup>6</sup> was cited as a good example of a virtual globe with quality data.

The technology allows digital data to be served or viewed through an easy-to-use viewer and provide useful backdrops for mapping, but generally they prevent one from generating maps from the geobrowser. While experts in spatial technology might use virtual globes and geobrowsers extensively for business purposes, the general public use them primarily for entertainment purposes. With virtual globes and geobrowsers, the business opportunities are not limited to the lack of data. One respondent felt that virtual globes/geobrowsers were having no impact on official mapping and several respondents did not know if they were having an impact.

### 10. Impacts of virtual globes/geobrowsers on official mapping through to end of 2014

The responses varied more than did those from CODIST, but were also generally

2004]. For example, in my presentation at the GISSA meeting [Cooper 2009b], I gave the example of VGI contributed on Google Earth, showing what was claimed to be pirate boats on the beach at Eyl in Somalia [“expedition” 2009] — the boats might then be at sea when the updated image is loaded on Google Earth and the KML would then point to an empty beach. See Figure 6.5.

<sup>6</sup>Did they mean Tracks4Africa?

## 7. Perceptions of virtual globes, VGI and SDIs

positive. Several felt they would improve the quality of the data and maps because there will be more pressure to supply accurate and up-to-date data as the demand increases and because people with access to geobrowsers will become more critical of map updates (though they are currently in a minority in South Africa). Other responses were that they will drive the priorities or initiatives of official mapping; will provide easy delivery of map updates; will enhance knowledge of 'where'; will result in the virtual obsolescence of paper maps; and will have a huge impact if their integration and use in education is done properly — or will have no impact because of the existing GIS awareness initiatives in the country!

One respondent hopes that virtual globes and geobrowsers will result in boundaries becoming standardized through a single entity, as the boundaries from various official organisations are not aligned, and that the postal code boundaries will be defined and made available. Virtual globes and geobrowsers may limit the need for GISc professionals and the quality of mapping may deteriorate as 'amateurs' feel they can do it themselves. Virtual globes and geobrowsers might not have a huge impact on information input because surveying companies supply government organisations with data. Again, several respondents did not know if they would have an impact over the next five years.

11. *Do you think that the legislative and policy environment in your country encourages or stifles innovation in the field of geographical information?*

A positive response, with none of the respondents feeling that legislation and policy stifle innovation with geographical information. Five felt that they encourage innovation and six felt that they neither encourage nor stifle innovation. Four had no opinion.

12. *Do you think that the legislative and policy environment in your country encourages or stifles the development of spatial data infrastructures (SDIs)?*

Again, none of the respondents felt that legislation and policy stifle SDIs — unsurprising, given the SDI Act [South Africa 2003]. Six felt that they encourage SDIs, three felt neither and two had no opinion.

13. *Do you think that the legislative and policy environment in your country encourages or stifles the development of, use of, and adherence to, standards?*

As with the CODIST responses, a fairly positive response, with seven replying that legislation and policy encourage standards, two replying neither, two replying stifle, and one did not know.

14. *How well do you think the legislative and policy environment in your country deals with issues such as virtual globes, volunteered geographical information and open access to geographical information?*

This resulted in a mixed response, with two selecting very well, one selecting adequately, four selecting poorly and one selecting not at all. Four did not know.

15. *Access to a virtual globe/geobrowser at home*

Eleven respondents have access to a virtual globe/geobrowser at home and one does not.

## 7. Perceptions of virtual globes, VGI and SDIs

### 16. Access to a virtual globe/geobrowser at work

Again, eleven respondents have access to a virtual globe/geobrowser at work and one does not (but not the same respondent as in the previous question). Hence, everyone who got this far through the questionnaire has access to a virtual globe and geobrowser at home or at work, or both. All of the respondents probably live and work in the metropolitan areas of Gauteng where connectivity is generally reasonable, though expensive.

### 17. Use of a virtual globe/geobrowser for personal purposes

Ten of the respondents use a virtual globe/geobrowser for personal purposes, and two do not.

### 18. Use of a virtual globe/geobrowser for work purposes

Eight of the respondents use a virtual globe/geobrowser for work purposes, and four do not. Overall, eleven use a virtual globe/geobrowser and one does not.

### 19. The virtual globes/geobrowsers used

All twelve of the respondents who answered use Google Earth. Seven also use others: four respondents also use NASA World Wind, one also uses Open Street Map, another also uses Microsoft Virtual Earth, and two also use Yahoo! Maps. Other virtual globes/geobrowsers that are used are ArcGIS Explorer (by two respondents), Tshwane street map guide, Open GIS and Global Mapper. As with the CODIST responses, this is also an indication of the dominance of Google Earth. However, as these GISSA respondents are heavier users of the technology, it is unsurprising that they have explored and used the alternatives.

### 20. The main reasons for using a virtual globe/geobrowser

The respondents could select several options if they so chose, with the results being as follows in Table 7.2:

Table 7.2: Reasons for using a virtual globe/geobrowser (GISSA respondents)

Reasons	Responses	Comments
Travel planning (work or leisure)	7	Unsurprisingly popular, in my opinion.
Providing a geographical context to news items	3	Again, the low response might be an indication of limited bandwidth, in that a user would not use a news Web site and a virtual globe simultaneously.
Accessing data for work purposes	5	Again, this option might have been misunderstood.
General curiosity	6	Again, unsurprisingly, a common activity.

*Continued on next page*

## 7. Perceptions of virtual globes, VGI and SDIs

Reasons	Responses	Comments
Publishing your data	3	This lowish response rate does not correspond well with the high response rates for using the virtual globe as a backdrop and for using a markup language in a geobrowser, which is surprising.
Reconnaissance for work purposes	7	Unlike the CODIST responses, with more power users this probably does reflect the use of the data on virtual globes for planning work activities.
Providing a geographical context to correspondence from friends and family	4	The lowish response rate correlates well with that for <i>providing a geographical context to news items</i> .
Backdrop for other geographical data	8	This high response rate correlates well with the high number of users of a markup language in a geobrowser.
Armchair travelling	7	Quite a different response rate from CODIST!
Searching for data	5	Again, with more power users one would expect more respondents to be using virtual globes to search for data sets they can download.
Other (please specify)	4	Quite a variety of other uses were provided here: research (could be covered by some of the uses listed, so it would be interesting to know what sort of research was envisaged by the respondent); Basic querying of data (again, several of the uses listed are really querying data); performing calculations of area and distance (not covered above, and there are other functions that geobrowsers provide); reviewing data (a temporary form of <i>publishing your data?</i> ); and plotting the pilgrimage of a friend to allow their family and friends to track progress.

Eleven respondents selected options here, selecting at least two each. Eight selected at least four options each and a ninth respondent selected all the options and added three and “a lot more” under *other*. Clearly, an indication that this group includes power users of virtual globes and geobrowsers.

### 21. Use of VGI in a virtual globe/geobrowser

Of those who use a virtual globe/geobrowser, seven use VGI and three do not.

## 7. Perceptions of virtual globes, VGI and SDIs

---

With more respondents being heavy users one would expect a greater awareness of VGI — but these responses might also have been biased by the presentations and discussions at the meeting.

### 22. *Use of a markup language in a geobrowser*

Nine use a markup language in a geobrowser and one does not. Hence, more of these respondents use a markup language than use VGI: this would indicate that they are active contributors of data to virtual globes, supporting that they are power users.

## 7.5 Analysis of the results from CODIST-I and GISSA

In general, even though the response rate was low, there was much variety in the answers received, indicating quite disparate exposure to virtual globes, geobrowsers and VGI amongst the respondents. There were more power users amongst the GISSA respondents, who probably have ‘cheaper’ and ‘faster’ Internet access than many of the CODIST respondents. The power users are better informed about these technologies and data, as one would expect from their greater use of them. However, there appears to be a greater disparity within the GISSA respondents. The responses confirm previous research, but also raise questions that need further investigation.

The following confirms previous research, such as by Butler [2006], Goodchild [2007b] and Sui [2008].

- Virtual globes and VGI promote geographical information in general, but they are perceived as threats to commercial GISs and to official mapping.
- Virtual globes provide quick and easy access to free data, the ability to share data and require low skills to access the data.
- Virtual globes and VGI encourage democratization (broader participation) by allowing ordinary people to contribute data quickly and easily that are then globally available — the wisdom of the crowd.
- A key concern, evident from the responses, is the legitimacy, quality, veracity and persistence of VGI. The quality is perceived to be quite variable, while the requirement is that data need to be up to date and accurate. See Section 6.7.1 for a discussion on the dimensions of quality and VGI.
- Similarly, a key concern is the inadequate nature of the available metadata for VGI and the (perceived) lack of moderation and verification. McDougall [2010] considers the quality of VGI to be the most contentious issue and other sources confirm the quality and metadata concerns [Goodchild 2007b; Craglia *et al* 2008]. See Section 5.12 for a discussion on metadata and VGI.
- Another concern is that naïve users can place too much faith in the reliability and accuracy of VGI. Goodchild [2007b] contemplated whether VGI, which relies on the essential ‘goodness’ of people in the virtual community, will in future be subjected

## 7. Perceptions of virtual globes, VGI and SDIs

---

to antisocial elements, much as the early days of the Internet were characterized by a certain altruism that was later ‘invaded’ by spam, viruses, and denial-of-service attacks. This concern is confirmed by the false reports about the Haiti earthquake [Palmer 2010], for example, wherein a bridge in Japan destroyed by an earthquake in 2006 was claimed to be a bridge destroyed by the earthquake in Haiti in 2010.

- There are concerns over bias in VGI, which are highlighted in the studies that attempt to understand the motivation behind VGI contributions [Budhathoki *et al* 2009; Coleman *et al* 2009], see Sections 8.4 and 9.6.
- Further concerns relate to transgressing privacy (as surveillance is now available to anyone), the security of the VGI, the exposure of sensitive sites and the use of VGI by vandals and criminals. These concerns have been confirmed by the likes of Goodchild [2007*b*] and Sui [2008].
- The respondents consider Google Earth to be the dominant virtual globe. See Section 8.3 for a discussion on Google Earth and other repositories of VGI.
- The respondents have diverse uses for virtual globes, particularly general curiosity and reconnaissance for work purposes. Other common uses are travel planning, accessing data for work purposes, using them as a backdrop for other geographical data, and armchair travelling.
- There is a moderate use of VGI in virtual globes by the respondents, and a low use of markup languages in a virtual globe by the CODIST respondents, but a high use by the GISSA respondents. The questionnaire did not attempt to gauge the intensity of the use of virtual globes.
- While VGI and virtual globes encourage democratization, one needs a computer, electricity and decent connectivity to be able to use a virtual globe, which respondents consider to be a problem. There is extensive use of mobile phones in Africa, even for accessing VGI, so this perception might be because the respondents themselves do not use VGI on their mobile phones. We consider research on the use of VGI contributions through mobile phones to be very important, especially in Africa, and have already embarked on further studies in this direction.

Below are issues that require further investigation:

- From the survey it is evident that virtual globes are having a limited impact on official mapping now (for example, which could be by forcing them to be more consumer-oriented), but they are expected to have a positive impact over the next five years, such as by encouraging better quality and improved availability of the data because of the competition from VGI.
- The legislative and policy environment is perceived to encourage the development of SDIs, and the development of, use of, and adherence to, standards, and to encourage more than stifle innovation in the field of geographical information.
- However, the legislative and policy environments deal poorly with issues such as virtual globes, VGI and open access to geographical information and require further research.

## 7. Perceptions of virtual globes, VGI and SDIs

---

### 7.6 Conclusions

Previous attempts have aimed at determining and categorising what motivates the contributors of volunteered geographical information, as discussed in Section 8.4. In contrast, this chapter reports on a survey to ascertain actual perceptions of VGI, virtual globes and spatial data infrastructures. A questionnaire was drafted to gather some data on the perceptions of these issues held by geographical information professionals from Africa, and the results have been reported here. These perceptions are important because they determine the future use of VGI and virtual globes in these communities.

This questionnaire has now been applied to two groups of GISc professionals with largely different backgrounds, experience with SDIs and access to virtual globes and geobrowsers. There was much variety in the answers received, indicating quite disparate exposure amongst the respondents. It would obviously be interesting to apply the questionnaire against other groups, such as GISc professionals in a country with cheap and abundant bandwidth, or the lay public in such a country. It would also be interesting to be able to apply the questionnaire to a sample that would provide a statistically meaningful representation of some population of interest. It might also be useful to update the questionnaire, addressing the weaknesses highlighted by the responses to date (eg: completion of only the first page and misinterpretation of the question on metadata), and making other appropriate changes. I intend to follow up the questionnaire with structured interviews with key people to improve the understanding of, for example, the intensity of use of virtual globes or the required legislative and policy environment for virtual globes, VGI and open access to geographical information in SDIs.

The results from the questionnaire have provided useful insights into the perceptions of geographical information professionals about virtual globes, VGI and SDIs. Some of the results confirm previous research, while others raise questions that warrant further research.

I would like to thank all the respondents for their willingness to complete the questionnaire.

### 7.7 Summary and looking ahead

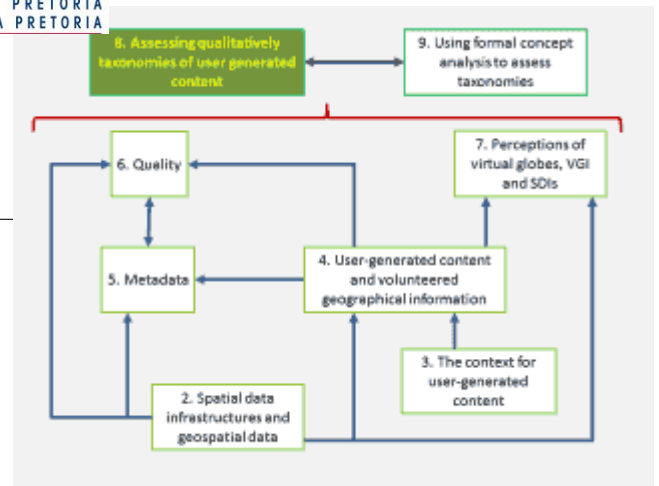
This chapter has reported on the results of a survey conducted through a questionnaire (included in Appendix A), of geographical information professionals in Africa in general and in South Africa in particular, concerning their perceptions of virtual globes, volunteered geographical information and spatial data infrastructures.

The next two chapters are the core of this thesis and build on the first five chapters. They provide assessments of both *taxonomies* of volunteered geographical information and *repositories* of volunteered geographical information that provide examples of the classes defined in these taxonomies. Chapter 8 presents a qualitative assessment of the repositories and taxonomies, both separately and against one another. Then, in Chapter 9, I present a more rigorous analysis of these taxonomies, using *formal concept analysis* (FCA).

---

*7. Perceptions of virtual globes, VGI and SDIs*

\*\*\*\*



## Chapter 8

# Assessing qualitatively taxonomies of user generated content

### 8.1 Overview of the chapter

In Chapter 7, I presented some perceptions that are held about volunteered geographical information. This is taken further in this chapter, where I present a qualitative assessment of various *repositories* of VGI and *taxonomies* of VGI, both separately and against one another. Then, in Chapter 9, I present a more rigorous analysis of these taxonomies, using *formal concept analysis* (FCA). This Chapter 8 discusses:

- The five taxonomies of UGC and VGI and the typology of citizen science, which are used in this chapter and in Chapter 9, see Section 8.2;
- The ten repositories containing VGI selected for assessing taxonomies, including their characteristics and possible candidates that were not selected, see Section 8.3;
- The qualitative assessment of the six taxonomies, using the ten repositories, and an assessment of them, see Section 8.4; and
- A preliminary taxonomy of user generated content that I developed, see Section 8.7.

Referring to Section 8.3, the quality in general of three of these VGI repositories is assessed in Chapter 6: SABAP2 in Section 6.8.1, OpenStreetMap in Section 6.8.2 and Tracks4Africa in Section 6.8.3. Further, please note that a preliminary version of the analysis in Section 8.4, together with the FCA analysis in Section 9.6, was published as a chapter, Cooper

---

## 8. Assessing qualitatively taxonomies of user generated content

---

*et al* [2012b], in the book “*Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications*” [Díaz *et al* 2012].)

The major original contribution that I have made that is presented in this chapter is:

- I assessed qualitatively several taxonomies of VGI against a selection of repositories of VGI. This showed that some of them could distinguish uniquely between the selected repositories, but not all. I then made some suggestions for improving these taxonomies.

Further, the key contributions that I have made that are presented in this chapter are:

- I have identified representative repositories that contain volunteered geographical information of various types and to varying extents. These are described in terms of why they are considered to contain VGI; their quality assurance (as this is considered to be a key weakness of VGI — see Section 6.2); whether or not they are crowd sourced (see Section 4.6); and the extent to which they contain professionally-generated content, as summarised in Table 8.2. These repositories are also used to assess five taxonomies of user-generated content and VGI, see Sections 8.4 and 9.6.
- I have assessed these taxonomies quantitatively, using these repositories to illustrate aspects.
- I have presented my own first attempt at a taxonomy of user-generated content (see Section 8.7), which was started before I assessed these other taxonomies, but which has been informed by my analysis of them. However, much work still needs to be done on this taxonomy, as it is incomplete and lacks definitions and examples (unfortunately, there are those who think that it is sufficient to use just the label to differentiate the classes). This taxonomy also needs peer review and should be subjected to the sorts of analysis done in Sections 8.4 and 9.6.

Broadly, the analyses presented in this chapter and the next have the objective of showing what is required for a robust taxonomy of user-generated content in general, and of volunteered geographical information in particular. This should aid the users, experts and researchers of user-generated content with identifying and understanding the content, as well as facilitating other theoretical research on user-generated content.

### 8.2 Taxonomies of UGC and VGI

A *taxonomy* has been described above in Section 2.4.1. Taxonomies are valuable because they enable a qualitative assessment of their membership. On the other hand, A bad taxonomy adds confusion.

User-generated content is described in Section 4.3 and volunteered geographical information in Section 4.5. As with many concepts introduced into Computer Science (eg: *ontology* and, for that matter, *taxonomy*!), *user generated content* has been interpreted in

## 8. Assessing qualitatively taxonomies of user generated content

different ways, whether because the concept is not well understood or because its ambiguity is being exploited, for whatever reason<sup>1</sup>. This then blurs the distinction between what is user generated content and what is not — that is, professionally generated content, or official content, or commercially generated content, or authoritative content, or even pseudo-authoritative content. The result is that one woman's user generated content could be another man's professionally generated content.

To clarify these issues for my research, I have made a first attempt at drafting a preliminary taxonomy of user generated content, which is presented in Section 8.7. Surprisingly, there appear to have been very few attempts yet at developing a taxonomy of user generated content, with the most comprehensive of those I found having been compiled by Coleman *et al* [2009] — coincidentally, also with a focus on volunteered geographical information. I have found the following:

- **Working Party on the Information Economy of the Organisation for Economic Co-operation and Development (OECD):** Wunsch-Vincent & Vickery [2007] presented various drivers of user-created content, and then included several other dimensions for classifying user-created content.
- **Gervais's taxonomy for copyright issues:** Gervais [2009] drew on the OECD taxonomy of user-created content, adding a dimension of taxonomy to cater for copyright issues.
- **Budhathoki, Nedovic-Budic and Bruce:** Budhathoki *et al* [2009] presented an overall framework for conceptualizing volunteered geographical information.
- **Coleman, Georgiadou and Labonte:** Coleman *et al* [2009] considered the nature and motivation of *producers* of volunteered geographical information.
- **Castelein, Grus, Crompvoets and Bregt:** Castelein *et al* [2010] characterized repositories of VGI (though their text implies they characterized VGI *per se*) from the perspective of SDI components, using the conceptual model of Rajabifard *et al* [2002].

These five classifications of user-generated content are discussed below in Section 8.4. Wiggins & Crowston [2011] developed a *typology of citizen science*, which is discussed in Section 4.4.2. Because some repositories of VGI can fulfil a variety of purposes, including contributing scientific data, the repositories are assessed against this typology in Section 8.5. Thereafter, my folksonomy is presented in Section 8.7.

In Section 9.6, the *discrimination adequacy* of the five published taxonomies has been assessed using *formal concept analysis (FCA)* [Priss 2006; Carpineto & Romano 2004]. Obviously, FCA is but one tool that can be used for analysing these taxonomies — their discrimination adequacy can also be determined manually. However, FCA provides a formal context for the analysis and the tools that support FCA, such as Concept Explorer [Yevtushenko *et al* 2003], facilitate such analysis. Indeed, using FCA has highlighted the paucity of VGI repositories with certain attributes and highlighted weaknesses with the published taxonomies, as discussed below in Section 9.6.

<sup>1</sup>For example, to publish a paper in a special issue of a journal.

## 8.3 Repositories containing VGI used for assessing taxonomies

### 8.3.1 Selecting repositories for analysis

To assess any taxonomy, one needs data that can be classified by the taxonomy: in the case of this thesis, different types of user-generated content, particularly those that constitute volunteered geographical information. However, it is difficult to characterise any particular set of UGC without having insight into the data set and its creation, or having comprehensive metadata available for the data set. Unfortunately, metadata for UGC is invariably sparse — or even non-existent (see Section 5.12). It is possible, instead, to create example data sets to classify according to the various taxonomies, though this may seem artificial.

Any repository containing UGC (whether made available over the Internet or not) could contain a wide variety of types of UGC (or even official content), depending on the explicit and implicit policies pertinent to the repository and the diligence with which they are applied. However, based on the review that I have conducted of repositories, it appears that the data in each of these repositories are sufficiently uniform that each repository can be characterised by its data and that they can also be discriminated from one another to a useful extent, using a taxonomy of UGC. The focus here is on VGI repositories specifically, rather than UGC repositories in general, because the focus of this thesis is on VGI.

Assessing repositories containing VGI is more useful than assessing just Web sites distributing VGI, as VGI repositories will include those accessed through Web sites, but will also include other VGI, such as might be distributed on CDs or DVDs. This is particularly relevant in Africa, where Internet access is often limited, very expensive and unreliable [6DISS 2005; Zennaro *et al* 2006], see Section 3.12. Also, strictly speaking, a virtual globe such as *Google Earth* [Google 2016a] is a *Web service* that provides access over the Internet to repositories of VGI and professionally generated content, but is not actually a *Web site*.

Further, a Web site is only a tool to access or display the data in a repository: one Web site can access multiple repositories and one repository can be accessed through different Web sites. Indeed, this is a reason for Web standards such as the Web Map Service (WMS), which produces static map images (that can be overlaid on one another) dynamically from available geospatial data [ISO 19128 2005]. Importantly, such Web services can produce results integrated from different sources and the services themselves can be *chained* (combined into a pipeline and processed *serially*) or *composed* (combined into a tree and processed *hierarchically*) [Rautenbach 2013; Rautenbach *et al* 2012a].

This section describes a selection of repositories containing VGI to varying extents, that will be used to assess the taxonomies introduced in Section 8.2. The assessment is in Sections 8.4 and 9.6, using qualitative assessment and formal concept analysis, respectively. The repositories are also assessed qualitatively against a typology of citizen science in Section 8.5.

Some of these repositories have also been included in the discussion on the types of VGI from the perspective of quality: see Section 6.9 and Figure 6.9. The repositories were

## 8. Assessing qualitatively taxonomies of user generated content

selected to provide a broad cross-section of the types of VGI repositories, while including those that would appear to be dominant and giving some preference to repositories based in South Africa. It is useful to present some sort of a popularity ranking of all of the candidate repositories, but too many of them carry too little traffic for the likes of Google [2016g] or Compete [2016], with the latter in any case having an explicit American bias. Alexa [2014]<sup>2</sup>, however, provides the rankings shown in Table 8.1, which are based on sites accessed from those users who have installed Alexa's tracking application. In some cases, the repository is a sub-domain whose ranking is not available on Alexa without a subscription, so the ranking of the parent site is shown, in brackets and italics.

Table 8.1: Rankings of repositories as at 2 April 2014, according to Alexa [2014]

Repository	Alexa rank, 2 April 2014	Alexa rank, 3 April 2015
Google <sup>3</sup>	( 1 )	( 1 )
Bing	21	24
Wikimapia	1 583	2 589
Tom Tom	4 301	3 365
OpenStreetMap	5 188	6 190
eBird	40 198	46 164
Ushahidi	135 665	180 781
Navteq	215 869	154 556
Bingmapsportal.com	248 101	289 062
FrontlineSMS	425 197	566 071
Tracks4Africa <sup>4</sup>	452 033	514 383
adu.org.za <sup>5</sup>	784 806	330 794
Cybertracker	1 768 998	1 279 247
Harassmap <sup>6</sup>	1 904 907	1 093 460
Mobilitate <sup>7</sup>	2 356 516	1 317 817
natworld.org (Natural World)	<i>Seems to be out of action</i>	
NASA Worldwind <sup>8</sup>		(1 056)
Yahoo! Maps <sup>9</sup>		( 5 )

*Continued on next page*

<sup>2</sup>Note that Alexa is owned by Google.

<sup>3</sup>Google Maps and Google Earth are not tracked separately, but on 2 April 2014, Google Maps received 1.55% of the visitors to google.com and Google Earth did not make the top 22 of Google sub-domains, getting 0.22% or less of the visitors to google.com.

<sup>4</sup>Ranked 6 730 in South Africa on 3 April 2015.

<sup>5</sup>SABAP2 is not a top-five keyword for searching for adu.org.za, being 3.52% or less of searches on 4 April 2014 and 3.95% or less on 3 April 2015.

<sup>6</sup>Ranked 29 052 in Egypt on 2 April 2014.

<sup>7</sup>No longer active.

<sup>8</sup>Worldwind is not a top-five keyword for searching for nasa.gov, being 1.1% or less of searches on 3 April 2015.

<sup>9</sup>Map is not a top-five keyword for searching for yahoo.com, being 1.29% or less of searches on 3 April 2015.

## 8. Assessing qualitatively taxonomies of user generated content

Table 8.1: Alexa rankings of repositories

Repository	Alexa rank, 2 April 2014	Alexa rank, 3 April 2015
Christmas Bird Count <sup>10</sup>		(27 002)
National Geographic's Field Expedition: Mongolia <sup>11</sup>		( 933 )
World Water Monitoring Day		4 844 800
Big Butterfly Count (UK)		<i>Too little data to be ranked</i>

Some repositories contain only official or commercial data that are professionally generated and some contain only user-generated content. However, given the nature of geospatial data and the prevalent use of base data to provide a context for, or to geocode, value-added data, many repositories that contain VGI also contain professionally-generated content to varying extents. Hence, there is a continuum from purely professionally-generated content through to purely amateur-generated content. Further, as discussed in Sections 4.3 and 4.8, the differences between professionally-generated and amateur-generated content are blurred and their quality can be similar.

### 8.3.2 The selected repositories

#### 8.3.2.1 Overview

As is clear from the illustrations of the data, as they were on 11 August 2011, for the area around the Nylsvley Nature Reserve in Limpopo, South Africa, (figures 8.1 – 8.4), there are significant differences between repositories containing VGI, even between those that could be expected to be similar: these repositories are all designed to contain base data and points of interest. These figures all show the default view for each repository's map for the area, without switching on or off any features, etc. For example, while Tracks4Africa (figure 8.1) had the most detail, especially points of interest, it lacked the railway line shown by OpenStreetMap (figure 8.2) and Wikimapia (figure 8.3). Bing Maps (figure 8.4) actually had terrain shading and more roads than the others, but these are not clear because of the default colour scheme. Bing Maps has been included in these illustrations because at the time in 2011, it included much VGI unique to it. However, the data in Bing Maps now come primarily from professional sources (national and commercial mapping agencies) and OpenStreetMap, so it has been excluded from the analysis below.

Please note that it is not the intention to make a value judgement of these repositories, but merely to illustrate how much variety there can be in repositories that are nominally similar. Hence, it also matters not that these images were taken a few years ago and that

<sup>10</sup>Christmas bird count is not a top-five keyword for searching for audubon.org, being 1.51% or less of searches on 3 April 2015.

<sup>11</sup>Mongolia is not a top-five keyword for searching for nationalgeographic.com, being 0.58% or less of searches on 3 April 2015.

## 8. Assessing qualitatively taxonomies of user generated content

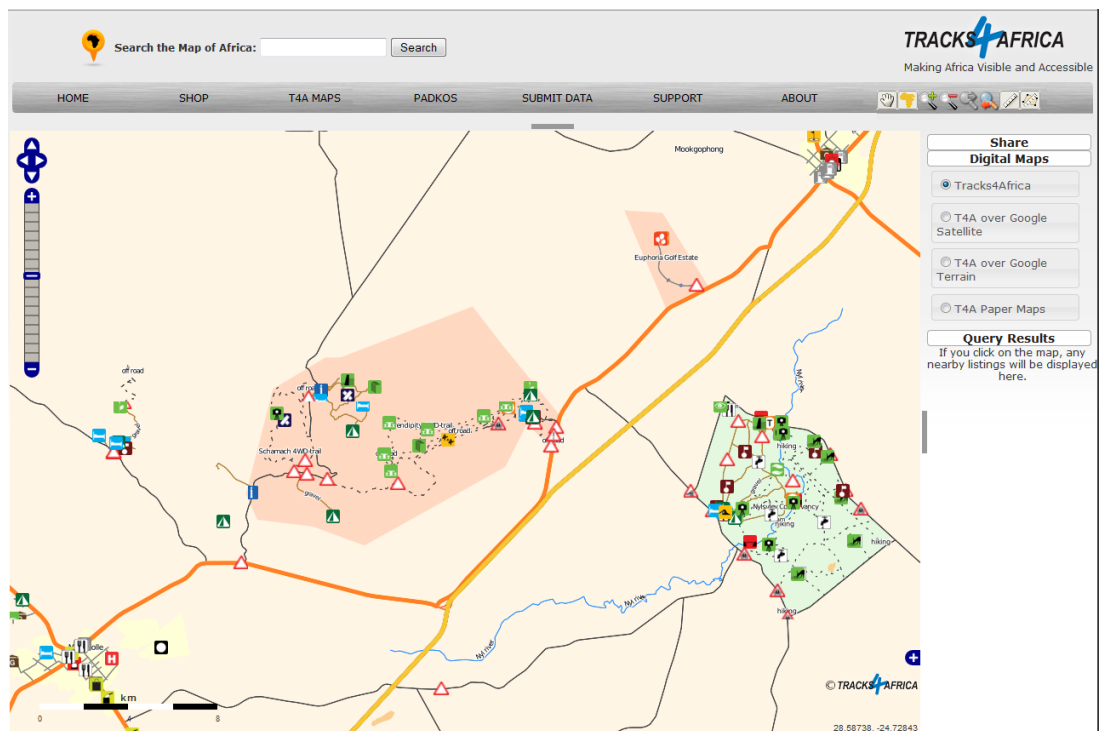


Figure 8.1: Tracks4Africa, showing the area around Nylsvley, Limpopo (11 August 2011) [Tracks4Africa 2016]

the data in the repositories have been updated since. Clearly, one could expect similar differences with commercial or official repositories, because all geospatial data sets are merely attempts to model and describe the world and each is always only just one of many possible ‘views’ of the world: see Section 5.2.

The repositories selected for assessing these taxonomies of VGI are Tracks4Africa, OpenStreetMap, Wikimapia, Google Earth, Google Maps, SABAP2, De Longueville *et al* [2010b], vehicle navigation, Mobilitate<sup>12</sup> and Harassmap. They are summarised in the following sections. Please note that the purpose is not to assess or criticise the repositories, but rather to use them to provide a variety of test cases for assessing the five classifications of VGI.

### 8.3.2.2 Tracks4Africa

*Tracks4Africa* is a repository and a Web site primarily of roads and tracks in Africa, but also of places of interest, see Figure 8.1. The network data are contributed in the form of GPS tracks, voluntarily and on their own initiative by individuals directly to the Web site, and hence are a classic form of VGI. Tracks4Africa synthesises the contributed data

<sup>12</sup>Please note that while Mobilitate no longer exists, it was still valid to use it for the analysis done in this thesis, as it represents a type of Web service and geospatial data repository. A similar service started recently in South Africa is LocalBlock [2016].

## 8. Assessing qualitatively taxonomies of user generated content

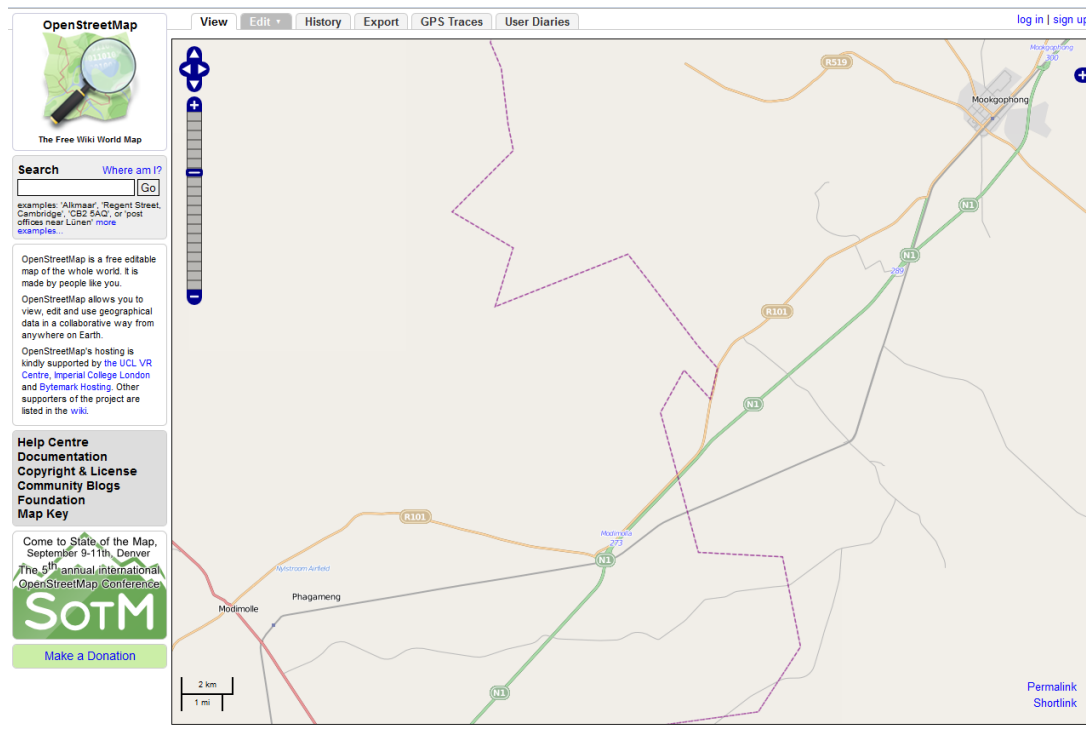


Figure 8.2: OpenStreetMap, showing the area around Nylsvley, Limpopo (11 August 2011) [OpenStreetMap 2016]

to produce the road networks, effectively using multiple entry for the quality assurance: synthesised data are only produced when they can corroborate the contributions of several different people. Only the synthesised data are made available, not the raw contributed data. The quality of Tracks4Africa in general is assessed above in Section 6.8.3.

Tracks4Africa also bundles in data to promote eco-tourism (eg: accommodation, fuel availability, and attractions), as obtained from its partner *Padkos* [Padkos 2016], and other sources. Tracks4Africa sells the data (updated twice a year), but the data are also available on Google Earth [Tracks4Africa 2016] and as part of the gROADS initiative<sup>13</sup> [gROADS 2014]. Tracks4Africa now also uses their data to make hardcopy maps, sold through retail outlets (eg: Tracks4Africa [2011]).

### 8.3.2.3 OpenStreetMap

*OpenStreetMap* is a repository and a Web site providing a free, editable map of the whole world, initiated as a repository of VGI, with *Wikipedia* [Wikimedia 2016] as its inspiration, see Figure 8.2. OpenStreetMap has a wide variety of tools for detecting possible errors (especially topological errors and missing tags) and procedures for publishing and

<sup>13</sup>Global Roads Open Access Data Set, Version 1, developed under the auspices of the Global Roads Data Development Task Group of CODATA, the Committee on Data for Science and Technology of the International Committee on Science (ICSU) [gROADS 2014].

## 8. Assessing qualitatively taxonomies of user generated content

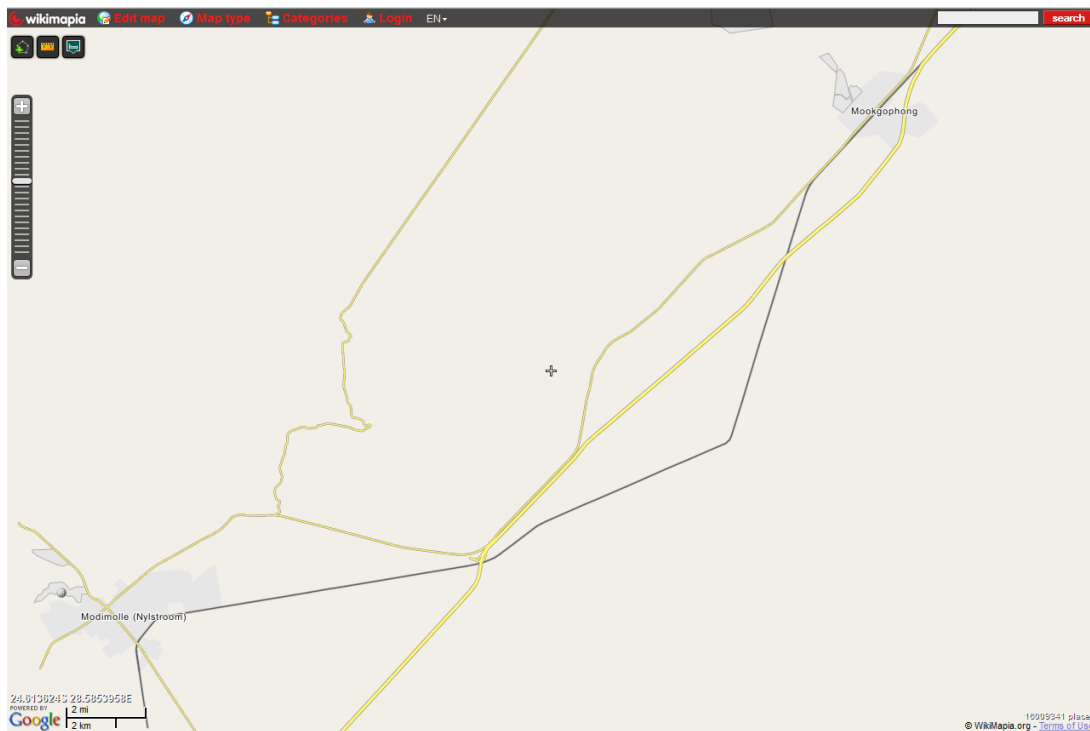


Figure 8.3: Wikimapia, showing the area around Nylsvley, Limpopo (11 August 2011) [Wikimapia 2016]

correcting detected errors. OpenStreetMap data are sometimes more up-to-date and of a higher quality than commercial or official data sets [OpenStreetMap 2016]. Several studies have been conducted to assess the quality of the data in OpenStreetMap, such as Haklay [2010]; Kounadi [2009]; Mooney *et al* [2010b]; Ather [2009]; Kounadi [2009]; Zielstra & Zipf [2010]; Govender [2011]; Du Plooy [2012]; Hankel [2012]; Camboim *et al* [2015]. The quality of OpenStreetMap in general is also assessed here in Section 6.8.2.

In collaboration with aid agencies, commercial satellite data providers and other organisations, OpenStreetMap made a significant contribution to mapping Port au Prince and other parts of Haiti for relief operations after the earthquake there on 14 January 2010 [Ball 2010a], for example. OpenStreetMap data are widely used and are available through other Web sites. OpenStreetMap also contains much data contributed by official mapping agencies.

While OSM has over one million registered members who contribute data, the bulk of the contributions or edits are done by a small proportion of the members, termed the “*senior mappers*” by Mooney & Corcoran [2014]. This correlates with the *long tail* exploited by many Web services [Anderson 2004], as discussed in Section 3.4.12. It appears that these senior mappers work primarily on their own [Mooney & Corcoran 2014; Du Plooy 2012].

## 8. Assessing qualitatively taxonomies of user generated content

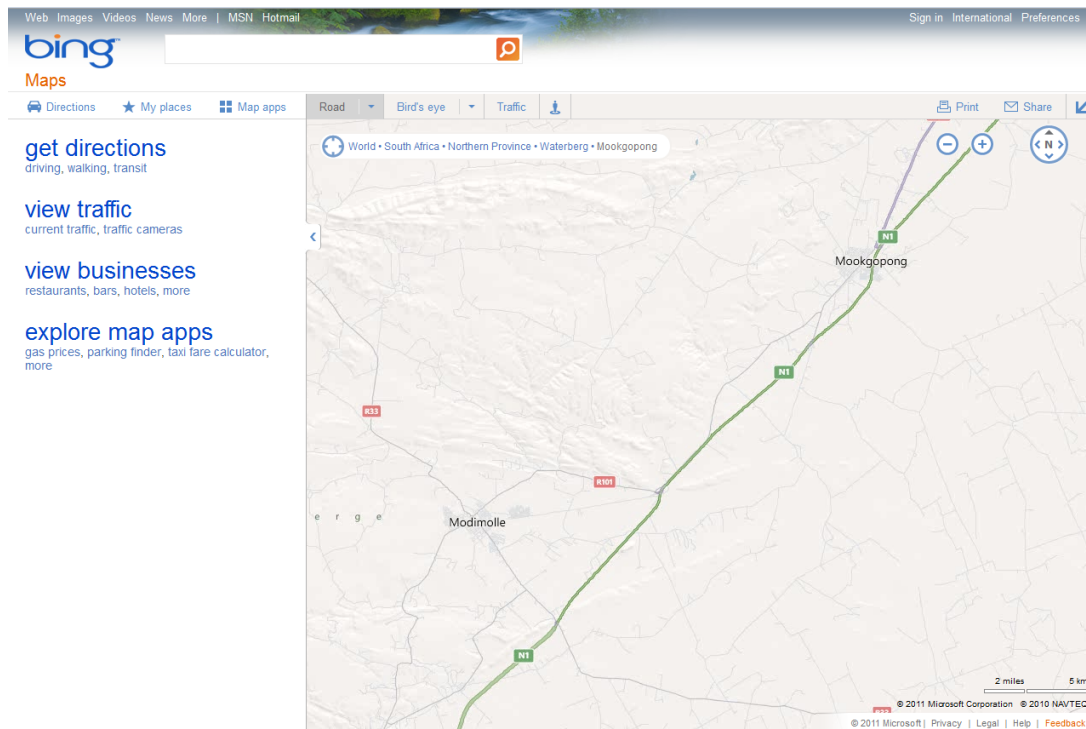


Figure 8.4: Bing Maps, showing the area around Nylsvley, Limpopo (11 August 2011) [Bing 2016]

### 8.3.2.4 Wikimapia

*Wikimapia* is a repository and a Web site providing a free, editable, interactive and multilingual map of the whole world, that uses Google Maps [Google 2016d] (maps, satellite imagery and/or terrain shading) as its default source of base data, see Figure 8.3. Wikimapia can also show photographs from Panoramio [Panoramio 2016] and provides a wiki that allows users to digitise polygons for adding geocoded information as notes. Interestingly, the Wikimapia interface also allows one to use the base data from OpenStreetMap, instead of from Google Maps. Wikimapia is not part of the Wikimedia Foundation, which hosts Wikipedia and related projects. There is a concern that some contributions are too subjective in describing places or for self promotion [Wikimedia 2016].

### 8.3.2.5 Google Earth

*Google Earth* is a virtual globe that is probably the best known, see Figure 8.5. It has also had a dramatic impact, both on making the lay public aware of digital geographical information and on promoting the creation and use of VGI [Perkins 2013; Harvey 2013a; Yu & Gong 2012]. Google Earth is not made available as a Web site (though it has a related Web site to download the client software (ie: the browser) and containing its conditions of service, etc), but as a Web service. The data on Google Earth are not directly

## 8. Assessing qualitatively taxonomies of user generated content

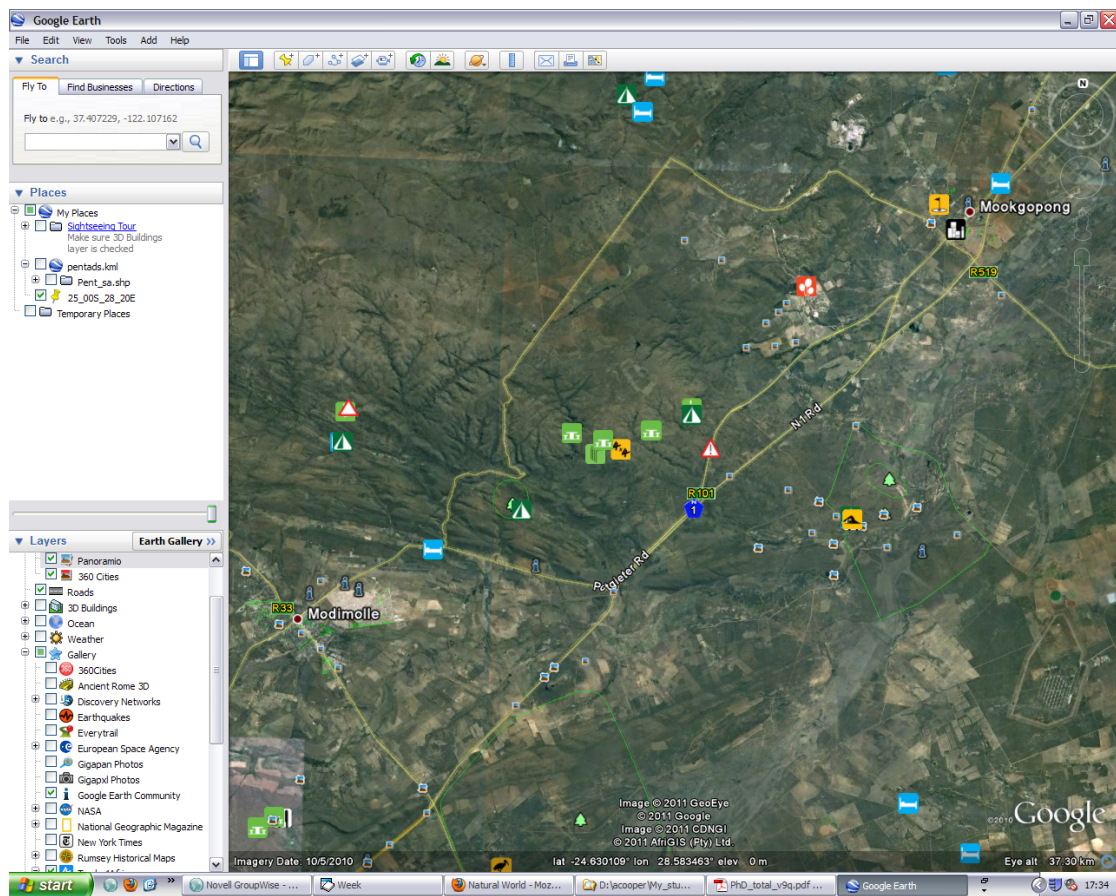


Figure 8.5: Google Earth [Google 2016a]

editable, but VGI can be contributed in bulk embedded in Keyhole Markup Language (KML) files, or through *Google Map Maker* (also used for *Google Maps*), geocoded photographs in *Panoramio* (started independently, but now owned by Google) [Panoramio 2016], or geocoded entries in *Wikipedia* [Google 2016a].

However, there have been complaints about Google's lethargy in removing spam generated by users through Google Map Maker, particularly spamming by service area businesses (SABs)<sup>14</sup> (eg: Austin [2012, 2013, 2014]; Shaw [2013b]). As a consequence, it is perhaps unsurprisingly that Google suspended editing through Map Maker from 12 May 2015, in response to obvious spam targeting a major corporation<sup>15</sup>, such as the depiction placed in Pakistan, of Google's Android robot urinating on an Apple logo [Kanakarajan 2015; Perez 2015; Siegal 2015].

<sup>14</sup> An SAB is a mobile service provider that operates at the user's location within a service area, rather than at the service provider's business premises. Examples of SABs are locksmiths, plumbers and "ambulance-chasers". It appears that it is common for such businesses to saturate Map Maker with false locations (even in areas where they are not licenced) to increase their chances of being high on a local search through Google [Austin 2014; Shaw 2013b].

<sup>15</sup> Which is much more likely than a small local business to be able to initiate a sustained legal action against Google.

## 8. Assessing qualitatively taxonomies of user generated content

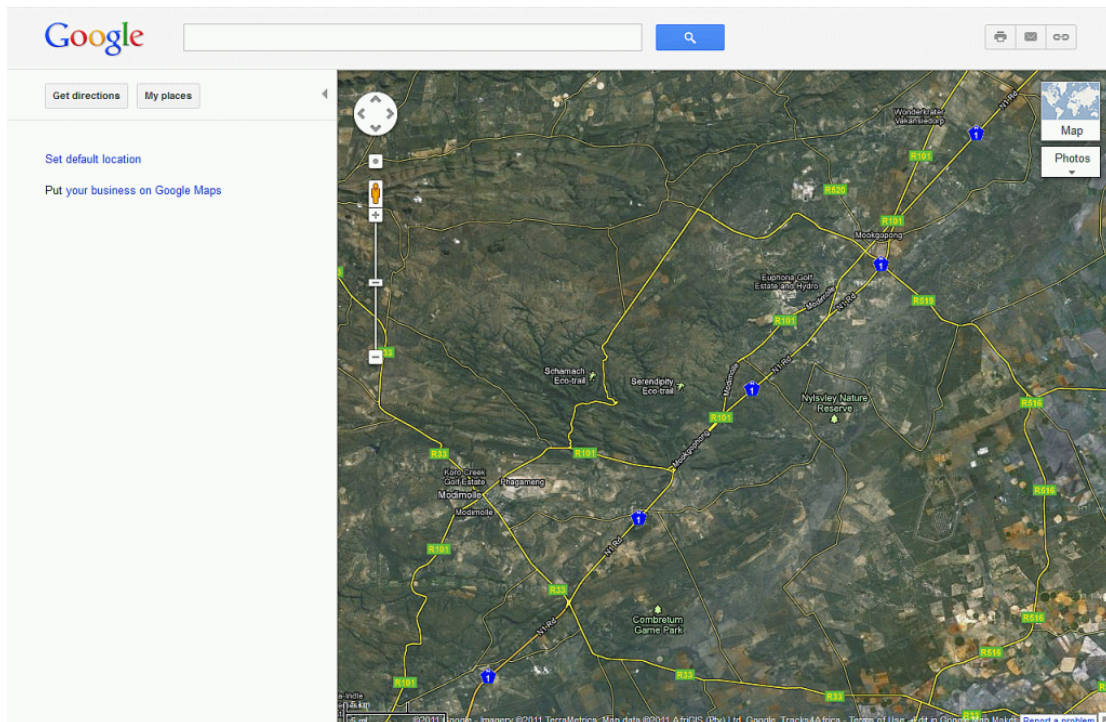


Figure 8.6: Google Maps [Google 2016d]

### 8.3.2.6 GoogleMaps

*Google Maps* is a Web site providing a free map of the whole world, see Figure 8.6, but not one that is directly editable. One can add VGI indirectly via geocoding entries in the likes of *Wikipedia* [Wikimedia 2016] and *Panoramio* [Panoramio 2016]. In addition, one can add and modify the listing of one's business on Google Maps [Google 2016d] and propose edits through Google Map Maker [Google 2016c]. However, as discussed above in Section 8.3.2.5, Map Maker has been suspended because it has been used to create spam. *Google My Maps* contains the VGI for Google Maps, but also allows a user to customise their interface to the data.

### 8.3.2.7 SABAP2

For the *Second South African Bird Atlas Project (SABAP2)*, begun in July 2007, volunteer bird watchers (and professional ornithologists) are gathering data according to a detailed, published protocol (recording bird distribution, observer effort and an index of abundance) and submitting the data either directly to the SABAP2 Web site, or by sending them through the post [Animal Demography Unit 2016b; Wright 2011], see Figure 8.7. As the data are gathered by geographical units, namely pentads (5' by 5'), and by temporal units (up to 5 days), the data are VGI — indeed, “the largest sponsorship for SABAP2 was contributed by the citizen scientists who participated in the project” [Underhill et al 2012]. There are also easy-to-use software tools available for capturing geocoded SABAP2 data

## 8. Assessing qualitatively taxonomies of user generated content

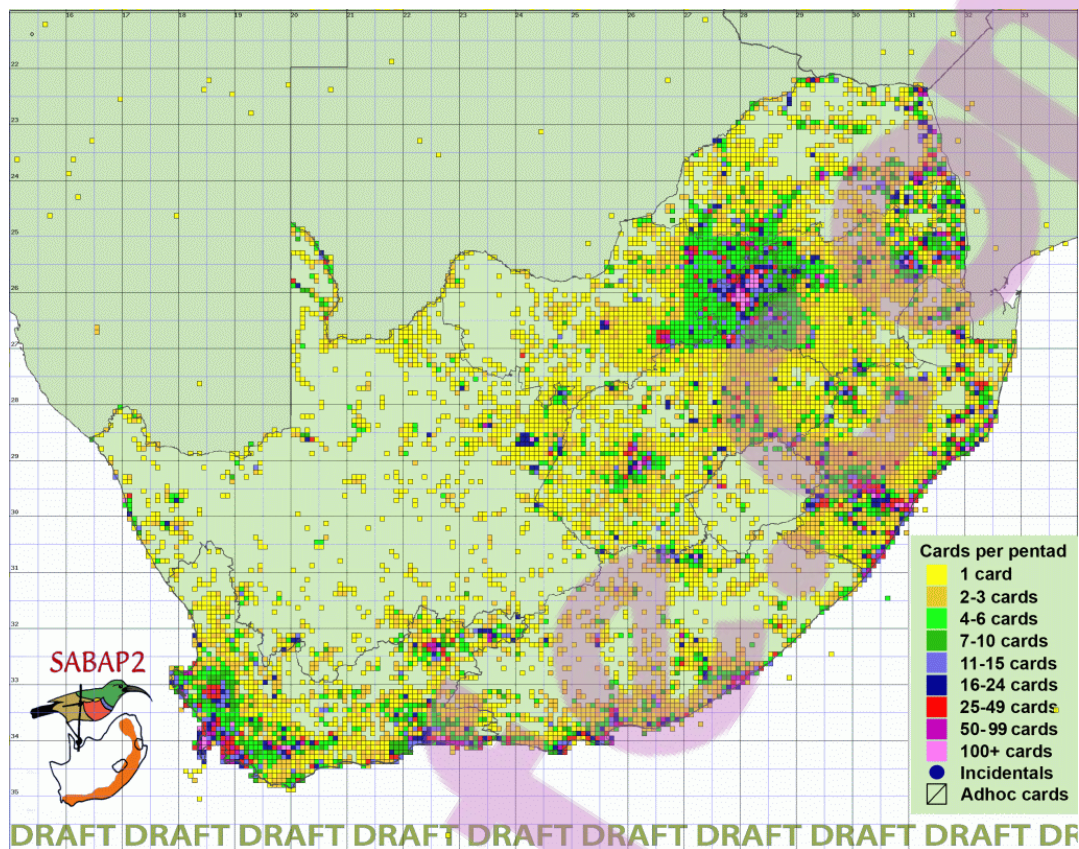


Figure 8.7: 2nd South African Bird Atlas Project (SABAP2) [Animal Demography Unit 2016b]

in the field and submitting one's records directly to SABAP2, using a mobile device with a GNSS receiver, such as CyberTracker [Liebenberg *et al* 1999; Liebenberg 2003; CyberTracker 2016] and Lynx Bird Ticks [2016]<sup>16</sup>.

SABAP2 is managed by the Animal Demography Unit (ADU) at the University of Cape Town. While the raw data are not made available on the Web site, various processed summary data sets are available and one can apply for the raw data. SABAP2 builds on SABAP1 (run from 1986 to 1997 [Harrison *et al* 1997, 2008; Underhill & Brooks 2014]), but with a more rigorous protocol with a finer spatial and temporal resolution that produces a better index of abundance. This allows for analysis that would not be possible with SABAP1 data, such as correlating the range expansion of the Common Myna *Acridotheres tristis* with human settlements [Underhill *et al* 2014]. See Bonnevie [2011]; Loftie-Eaton [2015] for issues with comparing SABAP1 and SABAP2 data, though. The quality of SABAP2 in general is assessed above in Section 6.8.1.

The same platform and similar protocols are being used for other atlassing projects, such as the butterfly atlas, the *Southern African Butterfly Conservation Assessment (SABCA)* [Animal Demography Unit 2016c], and the reptile atlas, the *Southern African Reptile Conser-*

<sup>16</sup>I have used both for SABAP2.

## 8. Assessing qualitatively taxonomies of user generated content

vation Assessment (SARCA) [Animal Demography Unit 2016d], but including them in this analysis as well would be duplication. Related citizen-science projects include the *Bird In Reserves Project (BIRP)*, the *Coordinated Waterbird Counts (CWAC)* and the *Coordinated Avifaunal Roadcounts (CAR)*, all also managed by the ADU. However, as SABAP2 is more comprehensive and as I know it better, it is assessed here, rather than the other projects. As of 17 January 2016, over 145 200 cards (field sheets) listing over 7 637 000 observations had been submitted by 1 853 citizen scientists to SABAP2, whereas over 32 300 cards listing over 1 648 000 observations had been submitted by 847 citizen scientists to BIRP, for example [Animal Demography Unit 2016b,a].

### 8.3.2.8 The perspective of De Longueville *et al* [2009]

*De Longueville, Ostländer and Keskitalo [2009]* have a different perspective, considering VGI to be data collected, synthesised and posted by the research team from interviews with stakeholders, many of whom, though not all, could be considered to be professionals and/or experts in the field (environmental data, in this case) [De Longueville *et al* 2010b]. The VGI component of their data was contributed by farmers and other residents in the area with extensive local knowledge, though because of the post-processing of the data by the project team, the VGI might well be tightly integrated with the professionally-contributed data. Hence, it is useful to include this alternative view of VGI in the analysis. The results from De Longueville *et al* [2010b] are not illustrated here because they did not provide an example in their paper.

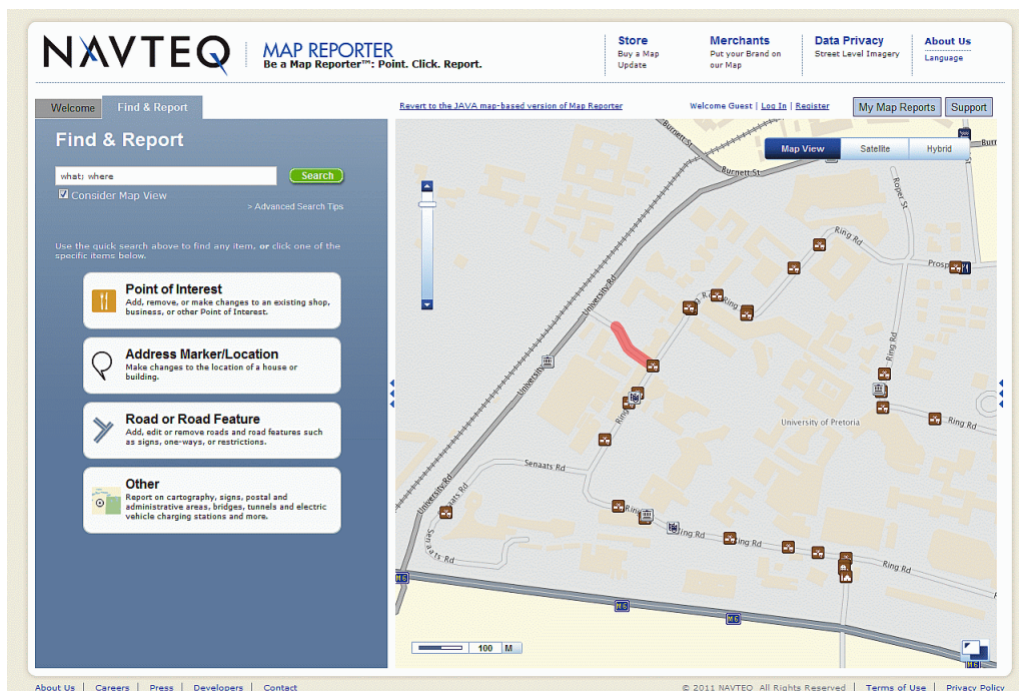


Figure 8.8: NAVTEQ's Map Reporter on 7 October 2011 [Navteq 2016]

## 8. Assessing qualitatively taxonomies of user generated content

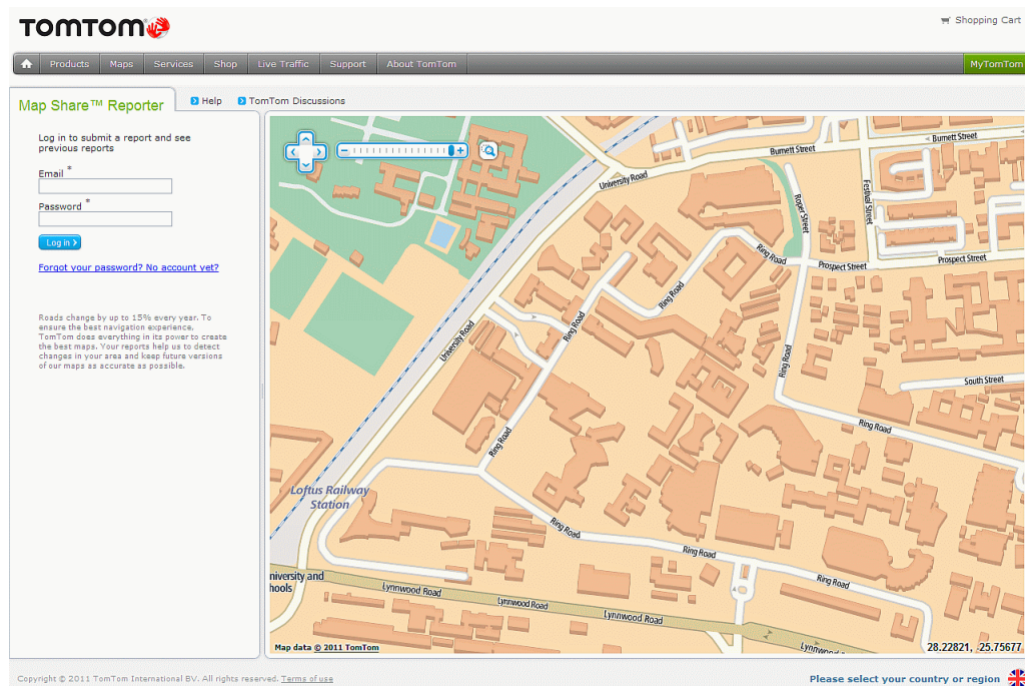


Figure 8.9: Tom Tom's Map Share on 7 October 2011 [TomTom 2016]

### 8.3.2.9 Vehicle navigation systems

Commercial *vehicle navigation systems* tend to provide their users with the capability of submitting corrections and updates to the map data (ie: contribute VGI), through either the in-car device or a Web site, see Figures 8.8 and 8.9, which both show the University of Pretoria on 7 October 2011. These updates can be distributed immediately to other users or put through a verification process first. Further, such systems might track their units remotely and anonymously (with or without informed consent: see Section 3.9.1 and Cooper *et al* [2009a] for a discussion of the ethical issues), and some consider such data to also be VGI. Examples of such systems are Tom Tom's *Map Share* [TomTom 2016], see Figure 8.9, and NAVTEQ's *Map Reporter* [Navteq 2016], see Figure 8.8. Please note that Map Reporter has now been included in HERE Map Creator [HERE Map Creator 2016].

While stand-alone portable tracking devices will probably be replaced by smartphones and built-in navigation systems in vehicles, the process and functionality remains the same. Further, the vendors of navigation systems are collaborating with those providing smartphones and built-in navigation systems, supplying hardware, software, data and services. Hence, they remain valid for this research.

### 8.3.2.10 Mobilitate

*Mobilitate* was a repository and a Web site that enabled citizens to use their mobile tele-



## 8. Assessing qualitatively taxonomies of user generated content

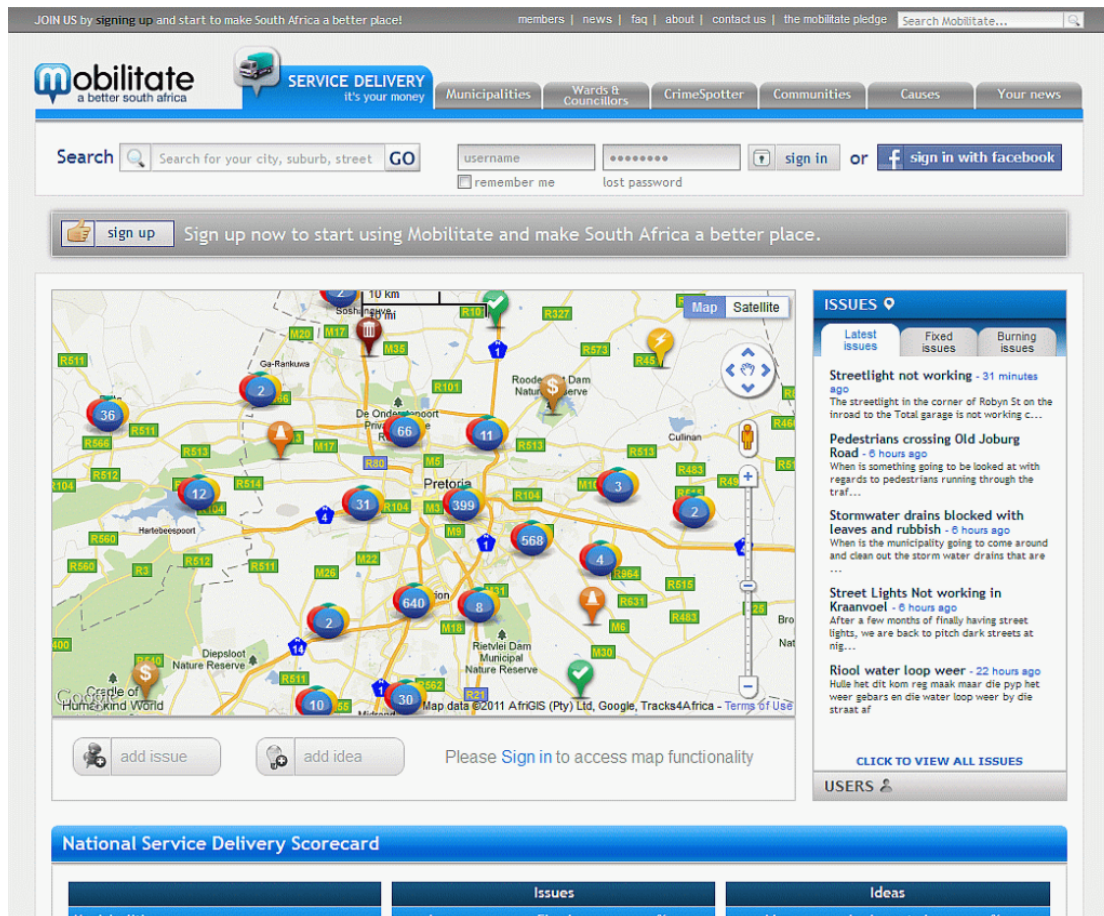


Figure 8.10: Mobilitate [Mobilitate 2015]

phones and Web browsers to log and prioritize public service problems in their own communities, and to communicate, organise, campaign, report and review, see Figure 8.10. It also provided local news and feedback mechanisms, involving suitably-motivated municipal councillors, formal structures such as community police forums (CPFs) and even metropolitan municipalities themselves (eg: issues reported on Mobilitate for the Cities of Cape Town and Johannesburg were fed through into the reporting systems of those metros) [Mobilitate 2015]. Many countries have similar Web sites, such as *Huduma — Fix my constituency!* in Kenya [Hudma 2016] and *SeeClickFix* [SeeClickFix 2016]. Unfortunately, it appears to have died during 2015. Nevertheless, it is useful to retain it in this analysis because it differs from the other repositories used in the analysis. A similar service started recently in South Africa is LocalBlock [2016].

### 8.3.2.11 HarassMap

*HarassMap* is a repository and a Web site for documenting sexual harassment of women (rape, touching, invitations, sexual comments, ogling, catcalling and facial expressions)

## 8. Assessing qualitatively taxonomies of user generated content

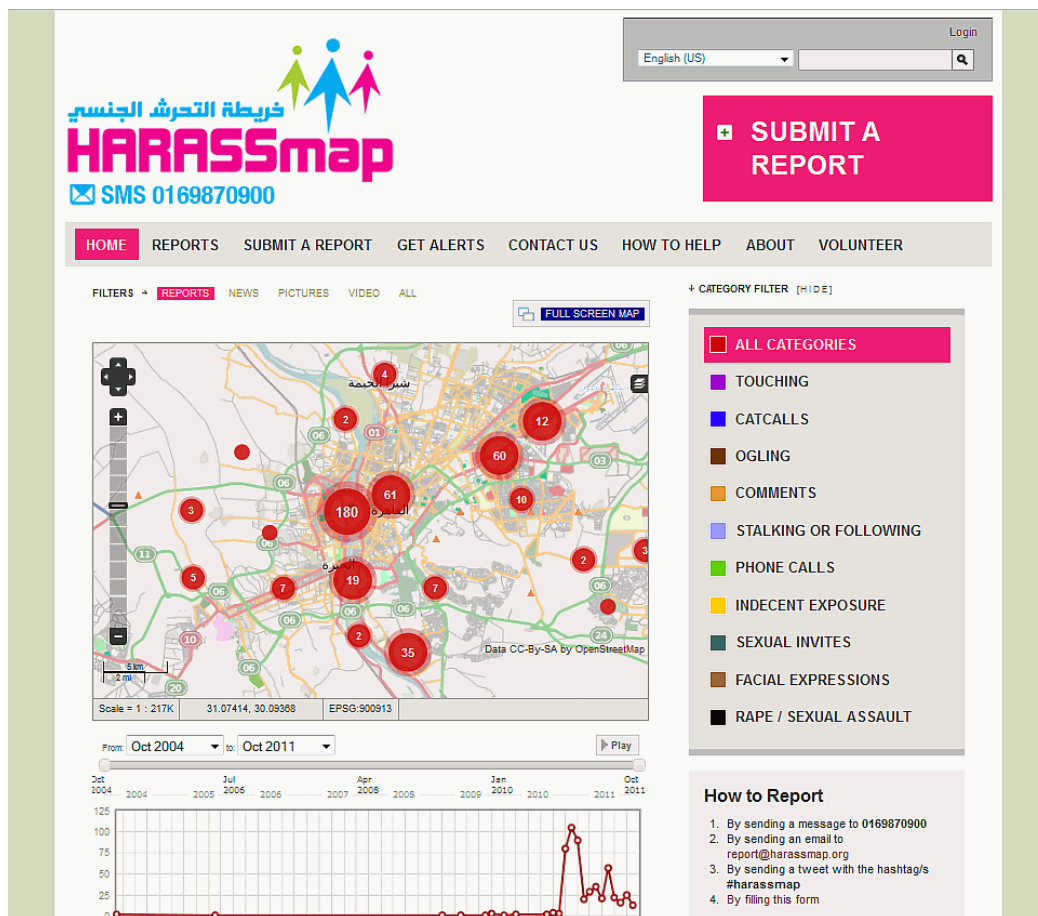


Figure 8.11: HarassMap [HarassMap 2016]

in Egypt, which is a very common problem there, see Figure 8.11. Reports are submitted via SMS, email, Twitter or Facebook. Some verification is done manually and some verification of the location is done by triangulating the source of the SMS [HarassMap 2016]. Harassmap uses the open-source platform *Ushahidi* [Ushahidi 2016], see Section 8.3.4. HarassMap's volunteers follow up on incidents to talk to those with a permanent presence in the street (doormen/women, shopkeepers, kiosk owners) to influence the culture on the street and to dispel the excuses for harassment and break stereotypes. They also run focus groups, conduct in-depth interviews and distribute literature. HarassMap has inspired similar initiatives that are up and running in eight other countries. Similar initiatives have arisen independently in other countries, such as *Chega de Fiu Fiu* ("Enough with the Catcalls") in Brazil [Diu 2015].

In their study on the effectiveness and use of HarassMap, Fahmy *et al* [2014] found that while the online map often obtained fuller and more comprehensive reports than the researchers got from interviews, it provided limited insights into differing definitions of harassment. It appears that the map provides a space where people are more willing to speak about the issue and with more anonymity. While the map is not perfect, they conclude that it is an effective tool for collecting data on sensitive issues [Fahmy *et al*

## 8. Assessing qualitatively taxonomies of user generated content

2014].

### 8.3.3 Characteristics of selected repositories

Table 8.2 provides a summary of the selected repositories of VGI that will be used below for assessing the VGI classifications. The characteristics described for each repository are:

- **Why VGI?**, which specifies why it is considered to contain VGI;
- **Quality assurance?**, which outlines what quality assurance is done on the data;
- **Crowd-sourced?**, which specifies whether or not it uses crowd sourcing; and
- **Professional content?**, which specifies what professional content is also included in the repository, if any.

Clearly, each of these characteristics could be used to produce a classification of VGI repositories, but that is beyond the scope of this thesis.

Table 8.2: Characteristics of repositories containing VGI

Repository	Why VGI?	Quality assurance?	Crowd-sourced?	Professional content?
<b>Tracks4Africa</b>	Purpose is to collect VGI; amateurs contribute their GNSS tracks	Multiple entry	Yes	Yes, including data from businesses wanting to be included in Padkos
<b>OpenStreetMap</b>	Purpose is to collect VGI; amateurs contribute anything	Suite of tools for detecting errors (especially topological errors and missing tags); procedures for publishing and correcting errors	In mapping parties and for special projects	Yes, provided by some mapping agencies
<b>Wikimapia</b>	Polygons of interest and updates	Partially through levels, but primarily through peer review and watch lists	Yes	Indirectly, as it uses Google Maps or OpenStreetMap for its base data

*Continued on next page*

## 8. Assessing qualitatively taxonomies of user generated content

Table 8.2: Characteristics of VGI repositories, *continued*

Repository	Why VGI?	Quality assurance?	Crowd-sourced?	Professional content?
<b>Google Earth</b>	Map Maker data, points of interest and geocoded Wikipedia and Panoramio, etc	None directly, but through Map Maker, Wikipedia, etc	Incidental	Yes, imagery and data provided by mapping agencies, etc
<b>Google Maps</b>	Map Maker data, geocoded Wikipedia and Panoramio data, and details of one's business	None directly, but through Map Maker, Wikipedia, etc, and self-promotion	Incidental, but done for North Korea specifically	Yes, imagery and data provided by mapping agencies, etc
<b>SABAP2</b>	Contributions from amateur bird watchers are essential	Automated checks based on SABAP1 and review by regional rarities committees	Yes, through periodic exercises, atlas bashes, etc	Professional ornithologists also contribute, but contributions are essentially indistinguishable
<b>De Longueville</b>	Gathered primarily from local farmers and others the authors consider are not domain experts	Post-processing by professionals	No	Some of the content is from professionals
<b>Vehicle navigation</b>	Users submit alerts and are often tracked	None on the reports, unless contradicted	Yes	Street networks are primarily professional content, supplemented by VGI
<b>Mobilitate</b>	Essential	Peer review or feedback from the local authority	Yes	Base data
<b>Harassmap</b>	Essential	Peer review and field work	Yes	Base data

### 8.3.4 Possible candidates that were not selected

The following virtual globes and similar Web sites were not included, because they do not carry VGI, or for other reasons.

## 8. Assessing qualitatively taxonomies of user generated content

- **Microsoft's Bing Maps.**

Bing Maps was included initially as it was a repository and a Web site providing a free, editable map of the whole world [Bing 2016], see Figure 8.4. However, it was excluded because the data in Bing Maps now comes primarily from professional sources (national and commercial mapping agencies) and OpenStreetMap. Bing Maps does allow for some VGI, as points of interest or Photosynth images, which are both catered for by several of the repositories analysed. Initially, the base data in Bing Maps were also unreliable. For example, as of 22 April 2010, for Tshwane many of the suburbs were displaced between 5 and 10 kilometres to the west and/or north relative to the road network, and the road network itself was both incomplete and inconsistent.

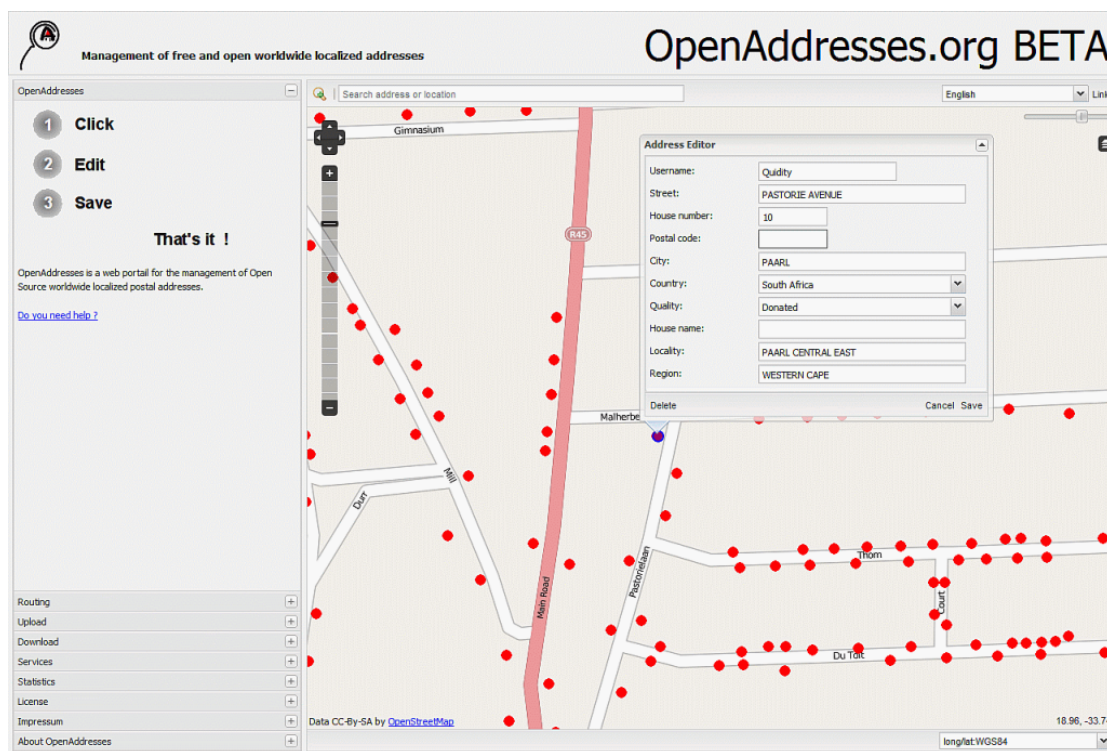


Figure 8.12: OpenAddresses, showing address data donated by Paarl (the red dots) [OpenAddresses 2016]

- **OpenAddresses.**

OpenAddresses was a repository and a Web site providing free, editable postal addresses, though they were actually primarily street addresses [OpenAddresses 2016; Stark 2011, 2012]. Many towns provided their address data to OpenAddresses, including Paarl and Wellington in South Africa, but many amateurs also contributed address data as VGI. OpenAddresses used OpenStreetMap for its back-drop. Figure 8.12 shows OpenAddress for part of Paarl, showing the official address data. Unfortunately, OpenAddresses closed down because of the lack of resources for maintaining the system and because it was not possible to incorporate OpenAddresses into OpenStreetMap [H-J Stark 2015, *pers comm*].

## 8. Assessing qualitatively taxonomies of user generated content

- **NaturalWorld.**

*NaturalWorld* was a repository and a Web site primarily for storing and making available records of sightings of species, especially birds, either by pentad (as for SABAP2) or by coordinates [NaturalWorld 2016], see Figure 8.13. Some SABAP2 contributors also submitted their species records directly to NaturalWorld, particularly as NaturalWorld allowed for the uploading directly of records from *CyberTracker* [CyberTracker 2016; Liebenberg *et al* 1999; Liebenberg 2003]. NaturalWorld also included historic data and data from ornithologists. However, it is no longer available.

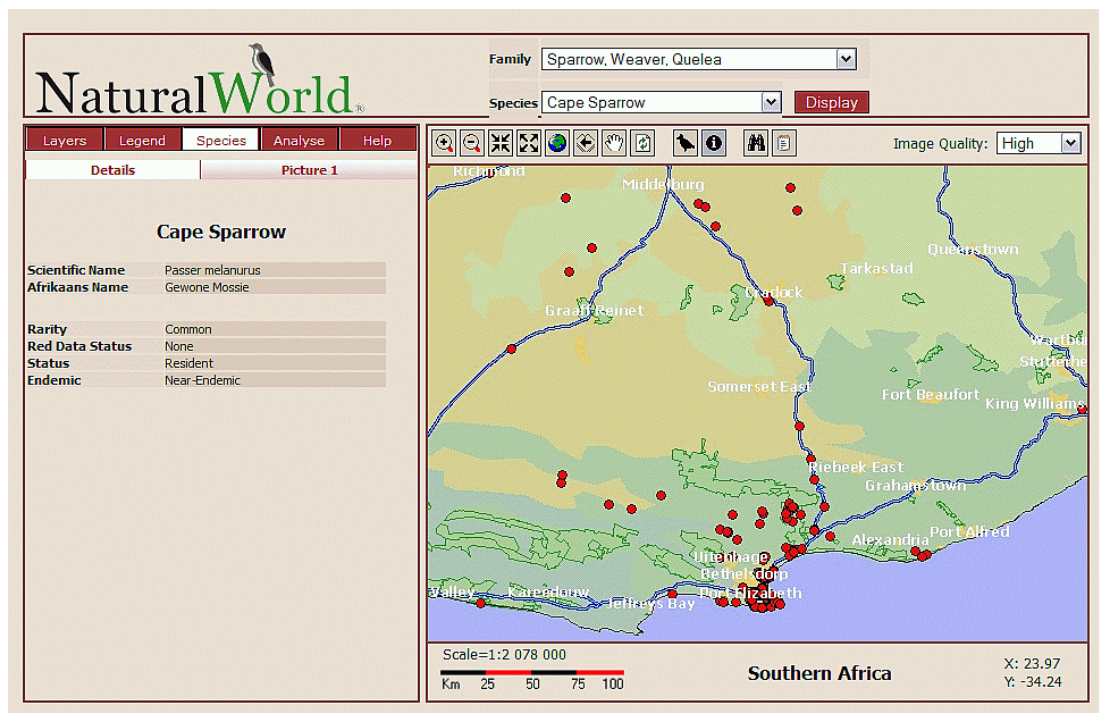


Figure 8.13: NaturalWorld [NaturalWorld 2016]

- **NASA World Wind.**

World Wind is an open-source, cross-platform API (application program interface) and SDK (software development kit) for a 3D virtual globe. That is, it is not a completed application targeted at end users, such as Google Earth, but is a set of tools for others to build their own applications, such as monitoring weather patterns, tracking vessels and analysing data. While it carries much interesting data (produced primarily by NASA and the USGS, it would appear), one could not add VGI initially [NASA 2016]. However, NASA has recently introduced a programme for VGI generated from its imagery and for adding other data sets to World Wind for general use [Goworldwind 2016].

- **Yahoo! Maps.**

It appears that while one might be able to geocode data on other systems, such as *Flickr* [Flickr 2016], that will then be presented on Yahoo! Maps, one cannot add

## 8. Assessing qualitatively taxonomies of user generated content

VGI directly on Yahoo! Maps [Yahoo! 2016].

- **Ushahidi.**

This is a collaborative project building the *Ushahidi Platform* and related open source software for information collection, visualization and interactive mapping. It began as a Web site to map reports of the post-election violence in Kenya early in 2008, evolved into tools for gathering crisis information from the public at large in near real-time, and is now applied more widely to build repositories of VGI and related services [Ushahidi 2016]. Generally, *Ushahidi* is used with tools such as *FrontlineSMS*, which provides a two-way group messaging hub for SMS-based messaging [FrontlineSMS 2016], with the texts being manually geocoded for inclusion in the application running on Ushahidi. Unfortunately, Camponovo & Freunds Schuh [2014] found high error rates in the classification of messages by volunteers, during the Haiti earthquake crisis in January 2010. Hence, as it is not a repository itself, *Ushahidi* has not been included in this analysis.

While the *Ushahidi* Web site does host some VGI repositories, they tend to be short-term projects, such as to monitor an election or respond to a crisis (eg: *Syrian Spring 2011*). In September 2011, the British Geological Survey started using *Ushahidi* for a citizen-science project, to collect geological observations [BGS 2011]. I felt that it would be more useful to include a repository built on Ushahidi with more permanence and a reasonable amount of data, which is why *HarassMap* was included instead (see above). Meier [2012] describes some applications of Ushahidi.

- **Other Citizen-science projects.**

There are many citizen science projects around the world, some of which are similar to SABAP2, see Section 4.4. The *Christmas Bird Count* in the United States of America is probably the longest running, having begun in 1900 [National Audubon Society 2016]. The *BBC Domesday Project* in the United Kingdom marked the 900th anniversary of the original Domesday Book in 1986 as a new survey of the country, with stories, photographs, videos and data provided by over 1 million people and stored on a multimedia GIS [BBC 1986; Rhind & Openshaw 1987]. A local and politically charged project is that of AfriForum to test the quality of potable water and treated sewerage in 125 municipalities [Odendaal 2014].

Other citizen science projects include *eBird*, for real-time online checklists of birds around the world [Cornell Lab of Ornithology 2016]; National Geographic's *Field Expedition: Mongolia*, for tagging clues in imagery for helping find the tomb of Genghis Khan [National Geographic 2016]; the *World Water Monitoring Day*, for gathering water quality parameters for local water bodies [WEF 2016]; *CitSci* for data on invasive species [Newman *et al* 2010]; and the *Big Butterfly Count*, which started in the United Kingdom in 2010 [Butterfly Conservation 2016]. However, while their protocols might be different, they are essentially the same as SABAP2 from the perspective of VGI taxonomies, so they have not been included as well.

## 8. Assessing qualitatively taxonomies of user generated content

### 8.4 Qualitative assessment of published taxonomies of UGC

#### 8.4.1 Overview of the taxonomies

Five published taxonomies of user-generated content are assessed qualitatively here: Wunsch-Vincent & Vickery [2007] in Section 8.4.2, Gervais [2009] in Section 8.4.3, Budhathoki *et al* [2009] in Section 8.4.4, Coleman *et al* [2009] in Section 8.4.5 and Castelein *et al* [2010] in Section 8.4.6. Only selections from these taxonomies are used in this section, as necessary to illustrate aspects of these taxonomies. The taxonomies are given in detail in Appendix B. The taxonomies are assessed using formal concept analysis in Sections 9.4 and 9.6.

#### 8.4.2 OECD Working Party on the Information Economy

As discussed in Section 4.3, there is no widely accepted definition of user-generated content, though Wunsch-Vincent & Vickery [2007] did provide a definition of *user-created content* (UCC), their term for user-generated content. Wunsch-Vincent & Vickery [2007] also present four groups of drivers of user-created content, which for reference, are given in detail in Section B.2.

- *Technological Drivers*: such as increased bandwidth, storage and processing speeds; better technologies and simpler software to create, distribute, and share content; better consumer devices for audio, photo and video; and the availability of UCC sites as outlets — and all of these at lower costs.
- *Social Drivers*: such as the shift to younger age groups (the *digital natives*) with the ICT skills and willingness to share content and even personal details; the desire to create and express oneself; the need for interactivity; the development of virtual communities and collaborative projects; and the spread of these social drivers throughout older age groups and their ability to fulfil societal functions, such as social engagement.
- *Economic Drivers*: such as the lower costs of many things, lower entry barriers and increased availability of tools creating, editing and hosting UCC content; increased possibilities for financing UCC ventures; increased interest of commercial entities to cater for UCC; the long-tail economics [Anderson 2004]; and better revenue streams through advertising and new business models.
- *Institutional and Legal Drivers*: such as schemes providing more flexible access to creative works; the right to create derivative works (eg: flexible licensing and copyright schemes such as Creative Commons); and end-user licensing agreements (EULAs) granting copyright to users for their content.

In terms of Maslow's hierarchy of needs [Maslow 1943] (see Section 3.8.16), the *technological, economic and institutional/legal drivers* help to make possible the *social drivers*, as they satisfy the lower "prepotencies". While the *social drivers* do still address some of the lower prepotencies (eg: when contributing to the likes of *Mobilitate* to improve the *safety*

## 8. Assessing qualitatively taxonomies of user generated content

of one's neighbourhood), they particularly address the highest prepotencies, *esteem* and *self-actualization*.

Clearly, many of the *technological*, *social* and *economic drivers* increase the viability of all the repositories (eg: bandwidth, willingness to share and lower costs), so the focus in Table 8.3 is on distinguishing drivers. In terms of the *institutional and legal drivers*, there are significant differences between the repositories, but users are often unaware of them because EULAs are too long and complex to read (see Section 3.9.1.4). As can be seen, there are two overlaps between the *technological* and *economic drivers*, with both including better software or tools for content and both including lower costs. It makes sense to include the better software only under *technological drivers* and the lower costs only under *economic drivers*. This has been done for Table 8.3, which shows the VGI taxonomies described in Section 8.3 assessed against this taxonomy from the OECD [Wunsch-Vincent & Vickery 2007].

Table 8.3: VGI repositories and UCC drivers [Wunsch-Vincent & Vickery 2007]

Repository	Technological	Social	Economic	Institutional/ legal
<b>Tracks4Africa</b>	Completely dependent on the availability of consumer GNSS receivers	Completely dependent on VGI	Completely dependent on willingness to share content and ease of uploading GNSS tracks	Contributors retain ownership of their content, but give Tracks4Africa licence to use the data as they wish. Tracks4Africa limits what may be done with their data
<b>OpenStreetMap</b>	Primarily dependent on the availability of consumer GNSS receivers	Primarily dependent on VGI	Primarily dependent on UCC tools and ease of uploading GNSS tracks	Dependent on flexible copyright
<b>Wikimapia</b>	Facilitated by better and technology and tools	Enhanced by VGI	Benefits from lower costs	Open content that can be shared, transformed and reused
<b>Google Earth</b>	Only possible because of the better and cheaper technology	Provides additional content and promotes it	Google's products thrive on advertising	Copying, modifying, reusing, etc is not permitted without a licence

*Continued on next page*

## 8. Assessing qualitatively taxonomies of user generated content

Table 8.3: VGI repositories and UCC drivers, *continued*

Repository	Technological	Social	Economic	Institutional/ legal
<b>Google Maps</b>	Facilitated by better and cheaper technology	Enhanced by VGI	Benefits from better tools and Google's products thrive on advertising	Copying, modifying, reusing, etc is not permitted without a licence
<b>SABAP2</b>	Helped by the availability of consumer GNSS receivers and base maps	Completely dependent on VGI	Largely dependent on ease of making contributions online	Helped by flexible licencing, allowing contributors to post their data elsewhere as well
<b>De Longueville</b>	Benefits from better and cheaper technologies	Dependent on willingness to share content	Probably benefits from better tools	Not specified
<b>In-car navigation</b>	Completely dependent on the availability of consumer GPS receivers and on bandwidth	Completely dependent on willingness to share content and personal details	Benefits from better tools	Copying, modifying, reusing, etc is not permitted without a licence
<b>Mobilitate</b>	Benefits from better and cheaper technologies	Completely dependent on willingness to share content	Largely dependent on ease of making contributions online	Content may be used for personal, non-commercial and information purposes only
<b>Harassmap</b>	Benefits from better and cheaper technologies	Depends on VGI and benefits from virtual communities and perceived anonymity of the Internet. Spreading to older people	Dependent on lower costs and better tools	Dependent on flexible copyright

Though not explicitly included in a taxonomy in their paper, Wunsch-Vincent & Vickery [2007] then describe various aspects of UCC/UGC, and from the section headings I have extracted the following taxonomy, as well. These would cut across their taxonomy of drivers, presented above.

- *Types of UCC*: Text, novel and poetry; photos/images; music and audio; video and film; citizen journalism; educational content; mobile content; and virtual content

## 8. Assessing qualitatively taxonomies of user generated content

[Wunsch-Vincent & Vickery 2007].

Broadly, all the repositories from Section 8.3 contain geospatial data and for each repository, their types of content are very varied, probably covering all possible types of UCC collectively — even novels, see Piatti *et al* [2009].

- *Distribution platforms*: Blogs; wikis and other text-based collaboration formats; sites allowing feedback on written works; group-based aggregation; podcasting; social network sites; virtual worlds; and content or file-sharing sites [Wunsch-Vincent & Vickery 2007].

The distribution platforms for the repositories are primarily Web sites and Web services, with most having accompanying blogs, wikis and presences on social media networks. Group-based aggregation is an essential part of all the repositories.

- *Monetisation of UCC and new business models*: Voluntary donations; charging viewers for services (pay-per-item or subscription); advertising-based models; licensing of content and technology to third parties; and selling goods and services to community [Wunsch-Vincent & Vickery 2007].

By definition, a VGI repository depends on voluntary donations, primarily of data, but also of funding and services. Some charge for data (eg: *Tracks4Africa*), and many exploit advertising, licencing of their content and technologies, and selling goods and services (eg: branded clothing). Some also get research funding, such as SABAP2 as a citizen-science project or *De Longueville et al* [2010b] as a research project.

- *Economic incentives along the value chain*: consumer electronics and ICT goods; software producers; ISPs and Web portals; UCC platforms and sites; users and creators; traditional media; professional content creators; search engines; Web services that profit from UCC; advertising; and marketing and brands [Wunsch-Vincent & Vickery 2007].

There is some overlap here with *monetisation of UCC content and new business models*, such as exploiting the brand of the repository, advertising and marketing. All the repositories provide services exploiting their UCC, such as data for scientific analysis (eg: SABAP2 and *De Longueville et al* [2010b]), real-time traffic densities (*in-car navigation systems*), promoting businesses to potential customers (eg: *Tracks4Africa* through *Padkos*, and *Google Maps*), and promoting the skills of the core teams responsible for the repositories.

- *Social impacts of UCC*: Changed information production leading to increased user autonomy, participation and communication; cultural impacts; citizenship engagement and politics; educational and informative impact; impact on ICT and other skills; and social and legal challenges of user-created content [Wunsch-Vincent & Vickery 2007].

Undoubtedly, *Google Earth* made a dramatic impact in promoting awareness and use of geospatial data when it burst onto the scene in 2005, but other VGI repositories have also had significant impacts. For example, in the *Map Kibera Project*, the

## 8. Assessing qualitatively taxonomies of user generated content

---

residents of Kibera, the massive slum in Nairobi, Kenya, have been producing the first maps of the area [Mulupi 2011], and as mentioned above, *OpenStreetMap* provided data for the relief operations after the 2010 earthquake in Haiti [Ball 2010a]. This led to the establishment of the Humanitarian OpenStreetMap Team (HOT), which provides geospatial data and other services for humanitarian responses and economic development, such as for the Ebola outbreak in West Africa in 2014 [HOT 2016].

- *Participative Web technologies*: Tagging; group rating and aggregation; syndication and aggregation of data; application mash-ups and open APIs; and file-sharing networks [Wunsch-Vincent & Vickery 2007].

Several of the repositories (eg: *Tracks4Africa* and *OpenStreetMap*) syndicate their data, such as to *Google Earth*. Several also encourage mash-ups and allow tagging of the data.

- *Digital content policies*: Enhancing R&D, innovation and technology in content, networks, software and new technologies; developing a competitive, non-discriminatory framework environment (ie: value chain and business model issues); enhancing the infrastructure (eg: technology for digital content delivery, standards and interoperability); business and regulatory environments that balance the interests of suppliers and users, in areas such as the protection of intellectual property rights and digital rights management, without disadvantaging innovative e-business models; governments as producers and users of content (eg: commercial re-use of public sector information); and conceptualisation, classification and measurement issues [Wunsch-Vincent & Vickery 2007]. Within the “business and regulatory environments” item under *Digital content policies*, Wunsch-Vincent & Vickery [2007] also provide a detailed classification concerning *intellectual property rights and user-created content*:
  - copyrights in the context of user-created content (original works created by users; derivative works; and facilitating UCC creation);
  - copyrights and the terms of services of UCC sites;
  - copyrights and the liability of UCC platforms;
  - digital rights management;
  - freedom of expression;
  - information and content quality;
  - mature, inappropriate, and illegal content<sup>17</sup>;
  - safety on the Internet and awareness raising;
  - privacy and identity theft;

---

<sup>17</sup>Of course, lumping “mature” (ie: “adult”) with “inappropriate” and “illegal” makes a judgement about “mature” content that is indefensible! For example, the first South African film to win the Academy Award for the Best Foreign Language Film of the Year was *Tsotsi*, which has an R (restricted) rating.

## 8. Assessing qualitatively taxonomies of user generated content

- impacts of intensive Internet use;
- network security and spam;
- virtual worlds, property rights and taxation;
- governments as producers and users of content.

Clearly, this aspect is wide ranging and overlaps with the other aspects, and particularly with the Institutional and Legal Drivers. As discussed below in Section 8.4.3, Gervais [2009] drew on this to develop an improved taxonomy for copyright issues. The digital content policies vary significantly across the repositories, with Google tending to claim ownership of everything added to *Google Earth* and *Google Maps* [Google terms of service 2010], to the Creative Commons Attribution-ShareAlike 2.0 license (CC BY-SA) [Creative Commons 2016] of *OpenStreetMap* and *Wikimapia*, allowing anyone to copy, distribute, transmit and adapt the data, as long as credit is given to the sources.

Wunsch-Vincent & Vickery [2007] felt that the conceptualisation, classification and measurement of digital content products and industries and their effects (statistical, economic and societal) are hard because of the lack of reliable, comparative data on UCC and changing usage habits.

### 8.4.3 Gervais' taxonomy for copyright issues

Gervais [2009] drew on the OECD taxonomy of user-created content [Wunsch-Vincent & Vickery 2007], identifying it as a matrix of type of content against distribution platform. However, even though it included 14 classes for *intellectual property rights and user-created content*, this did not meet his needs for understanding the copyright issues. So, he developed his own taxonomy of user-generated content for this purpose — it is important to realise that this taxonomy is independent of the presence or absence of any licences or contracts between the parties. While Gervais' taxonomy is obviously limited, it adds an important dimension

- *User-authored content.*

This is content authored without copying, derivation or adaption, and hence easy to deal with from the copyright perspective, as “the author is free to copy, upload, perform and/or make available” their content on any basis. A complication could arise when the author uses a Web site or a technology that takes a licence for the owner of the site or technology to use the content, as a condition of using the Web site. Typically, the end-user licence agreement (EULA) lets the user retain ownership while granting the owner of the site or technology broad rights to use, reproduce and transmit the content. Further, the EULAs often require the user to warrant that their contribution does not infringe anyone else's rights. The question then is to what extent is the content authored by the site or technology? For example, with *in-car navigation*, how should authorship be apportioned between the user being tracked passively (though creating the VGI) and the owner of the system? Presumably, this would need a law case to resolve.

## 8. Assessing qualitatively taxonomies of user generated content

- *User-derived content*:  
This is considered by Gervais to be the most complicated category, because of the nature of the underlying right and whether or not the derivation and/or reproduction constituted *fair use*, which is determined by the *use value* gained by the user and the *exchange value* lost by the rights holder. Examples of fair use for user-derived content include critiques and parodies.
- *User-copied content*:  
Merely copying pre-existing content is *prima facie* infringement and hence generally illegal and illegitimate. However, it could be considered to be *fair use* if only a “short excerpt” is used (determined qualitatively more so than just quantitatively) or if the use is “transformative”. A complication here is that the First Amendment [United States of America 1791] has been used in the United States of America as a defence, which is not necessarily applicable in other countries<sup>18</sup>. Examples of fair use for user-copied content are *framing* (including another Web site unaltered within a frame on one’s own Web site, without actually copying the content of the other Web site) and *thumbnail* images of Web pages for linking to them.
- *Peer-to-peer as UGC*: The key difference between this category and *user-copied content* would appear to be that *user-copied content* should be *transformative*, that is, that it does not “merely supersede the objects of the original creation” [US Court of Appeals for the Ninth Circuit 2007]. Gervais [2009] feels that while unauthorized peer-to-peer (P2P) file-sharing is generally illegal, it is not going away and controlled monetizing is the best outcome for both authors and users.

This taxonomy of Gervais [2009] is applied to the repositories in Table 8.4. Please note that I have classified the repositories according to the data they make available, rather than the inputs they receive. For example, while both Tracks4Africa and OpenStreetMap accept *user-authored content*, Tracks4Africa processes the data first and hence presents only *user-derived content*, while OpenStreetMap presents both. Fortunately, none of the repositories appear to contain *Peer-to-peer as UGC*.

Table 8.4: VGI repositories and copyright issues [Gervais 2009]

Repository	User-authored content	User-derived content	User-copied content	Peer-to-peer as UGC
Tracks4Africa		X		
OpenStreetMap	X	X	X	
Wikimapia	X			
Google Earth	X	X	X	
Google Maps	X			
SABAP2		X	X	
De Longueville		X		
In-car navigation	X	X	X	

*Continued on next page*

<sup>18</sup>Gervais is based in the USA and wrote from that perspective, though he was educated in Canada and formerly worked for both WIPO and WTO.

## 8. Assessing qualitatively taxonomies of user generated content

Table 8.4: VGI repositories and copyright issues, *continued*

Repository	User-authored content	User-derived content	User-copied content	Peer-to-peer as UGC
Mobilitate	X	X		
Harassmap	X	X		

### 8.4.4 Budhathoki, Nedovic-Budic and Bruce's framework for VGI

Budhathoki *et al* [2009] presented an overall framework for conceptualizing volunteered geographical information, but as this was a conference presentation, they did not define any of these terms, though they did provide a detailed expansion of one of their classes, *Motivation*, as discussed below.

- *Context of Participation*: personal, social or technological.
- *Motivation*: unique ethos, learning, career, personal enrichment, self-actualization, self-expression, self-image, self-gratification/fun, re-creation, social, group accomplishment, group attraction, group maintenance, identity, reputation, monetary, instrumentality, cognitive capital/self-efficacy, reciprocity, sense of community, meeting own need, freedom and creativity, altruism, trust in the underlying infrastructure, protective, structural capital, self-presentation, relation management, and socio-political motives.
- *Contribution Mechanisms*: structure, process or norms.
- *Contribution*: this is not explained, but presumably refers to the actual VGI provided.
- *Issues*: reliability, quality, value, privacy, copyright, coverage, credibility, sustainability and social justice.

These *issues* are largely subjective and to be determined by the user, not the producer.

For their class *Motivation*, Budhathoki *et al* [2009] provided 29 motivational factors, with conceptual definitions and literature sources for them. There appears to be overlaps between many of the motivational factors they list, particularly those related to self actualization, but it is not clear from their presentation material if they were merely documenting the motivational factors they had found in the literature, or if they were making value judgments on them.

Budhathoki *et al* [2009] also presented an analysis of the talk pages of *OpenStreetMap*, which effectively gives a taxonomy of the motivations of contributors to *OpenStreetMap*. Clearly, these will all also apply to the other generic VGI repositories (*Tracks4Africa*, *Wikimapia*, *Google Earth* and *Google Maps*), but also to *SABAP2* and *Mobilitate*, to varying extents. It is less likely that the informants for De Longueville *et al* [2010b] were motivated by *anti-corporate sentiment*, their *unique ethos* or the *visual power of a map*, but they might have had an *expectation of reciprocity*, in terms of what gets done in their area. Those contributing through *in-car navigation* are explicitly expressing a *pro-corporate sentiment*,

## 8. Assessing qualitatively taxonomies of user generated content

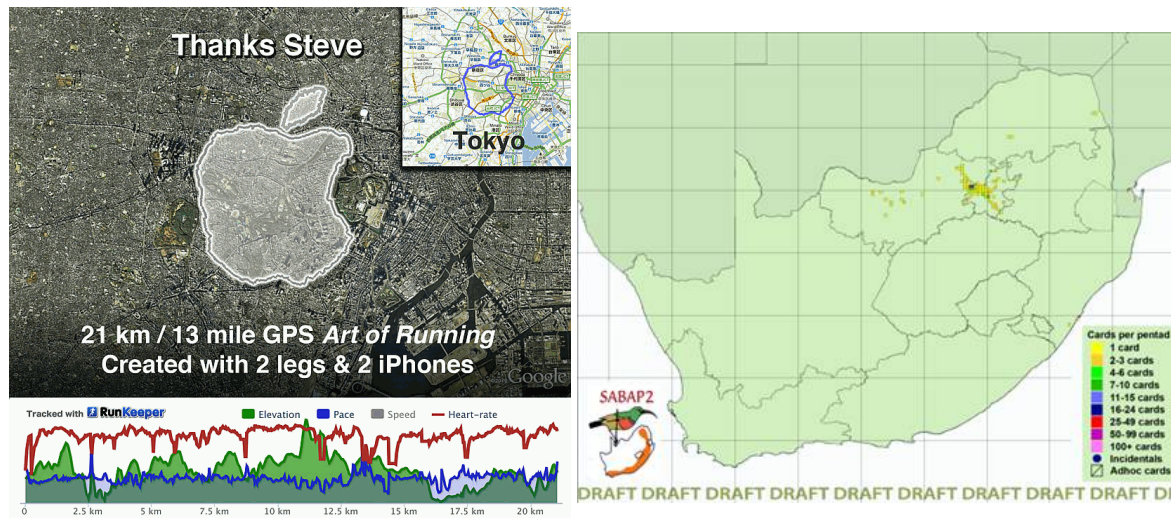
---

but some of the other motivations might apply to those contributors aware that they are being tracked or who provide traffic alerts. Surprisingly, Budhathoki *et al* [2009] did not include most of these explicitly in their list of *motivations* presented above, as indicated in my commentary on them below.

- *Fulfillment of self-need*:  
This is included above by a combination of *meeting own need* and others.
- *Anti-corporate sentiment (unique ethos)*:  
While Budhathoki *et al* [2009] included *unique ethos* above, this is not exactly the same as *anti-corporate sentiment*. One could contribute VGI to castigate a company randomly and on the spur of the moment (eg: map its pollution), rather than as part of any systematic expression of one's unique character or guiding principles. Similarly, one's unique guiding principles could be *pro-corporate* when contributing VGI, such as advertising a company's products, services and outlets.
- *Expectation of reciprocity*:  
This is included above, explicitly.
- *Visual power of a map (self-gratification)*:  
While *self-gratification* is included above, and manifests itself in the likes of GPS art or selecting the polygons to survey for a citizen-science project, the *visual power of the map* could go beyond that in highlighting or for highlighting significant spatial aspects or for political purposes. An example of GPS art is the half-marathon tribute to Steve Jobs, see Figure 8.14(a), [Tame, Joseph 2011], and an example of the latter is the contributions by Anthony Paton, an artist, who selected the pentads to map for SABAP2 to create the image of a bird, see Figure 8.14(b) [Retief 2011].
- *Outdoor activity (re-creation)*:  
Budhathoki *et al* [2009] define *re-creation* as “the process of forming anew or creating one's self again”, and hence it is **not** related to *outdoor activity*, which would be *recreation* defined as “activity done for enjoyment when one is not working” [Oxford 2016]. *Outdoor activity*, or more generically, this second meaning of *recreation*, has not been included in the *motivations* above.
- *Pride of local knowledge*:  
Not included above.
- *Concerns for a substantive issue (need)*:  
Not included above explicitly, though it is close to *socio-political motives*.
- *Other — explored: monetary, hobby, learning*:  
Both *monetary* and *learning* are included explicitly above, while *hobby* would be covered by *self-gratification/fun*, and possibly others.

This taxonomy of Budhathoki *et al* [2009] is applied to the repositories in Table 8.5. Clearly, there is a range of motivations for contributing VGI to each of the repositories.

## 8. Assessing qualitatively taxonomies of user generated content



(a) GPS art tribute to Steve Jobs [Tame, Joseph 2011]

(b) Pentad art work from SABAP2 [Retief 2011]

Figure 8.14: Examples of VGI used to create virtual land art

Table 8.5: VGI repositories and VGI framework [Budhathoki *et al* 2009]

Repository	Context of participation	Motivation	Contribution mechanisms	Contribution	Issues
Tracks4Africa	Personal, social	Unique ethos, self-expression, self-gratification/ fun, group accomplishment, reputation, reciprocity, sense of community, meeting own need altruism, trust in the underlying infrastructure	Norms	Tracks and points of interest	Reliable, quality, copyright, coverage
OpenStreetMap	Personal, social and technological (eg: software)	Unique ethos, learning, self-expression, self-gratification/ fun, social, group accomplishment, group maintenance, reputation, reciprocity, sense of community, meeting own need, freedom and creativity, altruism, trust in the underlying infrastructure, socio-political motives	Structure, process, norms	Anything	Reliability, quality, value, coverage, credibility, sustainability, social justice

Continued on next page

## 8. Assessing qualitatively taxonomies of user generated content

Table 8.5: VGI repositories and VGI framework, *continued*

Repository	Context of participation	Motivation	Contribution mechanisms	Contribution	Issues
<b>Wikimapia</b>	Personal, social	Self-expression, self-image, self-gratification/ fun, re-creation, social, identity, reciprocity, sense of community, meeting own need, altruism, relation management, socio-political motives	Structure, process	Points of interest	Value, coverage
<b>Google Earth</b>	Personal, social	Self-expression, self-image, self-gratification/ fun, identity, instrumentality, reciprocity, sense of community, altruism, trust in the underlying infrastructure, structural capital, self-presentation, socio-political motives	Structure	Anything, project data	Reliability, quality, value, copyright, coverage, credibility
<b>Google Maps</b>	Personal	Self-expression, self-gratification/ fun, identity, reputation, monetary, reciprocity, sense of community, meeting own need, altruism, trust in the underlying infrastructure, structural capital, self-presentation, socio-political motives	Process	Anything	Reliability, quality, value, copyright, coverage, credibility
<b>SABAP2</b>	Personal, social	Unique ethos, learning, personal enrichment, self-actualization, self-gratification/ fun, social, group accomplishment, reputation, instrumentality, cognitive capital/ self-efficacy, sense of community, altruism, trust in the underlying infrastructure, protective	Process, norms	Field sheets and individual sightings	Reliability, quality, coverage, credibility, sustainability

*Continued on next page*

## 8. Assessing qualitatively taxonomies of user generated content

Table 8.5: VGI repositories and VGI framework, *continued*

Repository	Context of participation	Motivation	Contribution mechanisms	Contribution	Issues
De Longueville	Social	Personal enrichment, social, group accomplishment, instrumentality, cognitive capital/ self-efficacy, sense of community, meeting own need, protective, structural capital, socio-political motives	Process	Gazetteer entries, testimonials	Value, credibility
In-car navigation	Personal	Learning, group attraction, instrumentality, meeting own need, trust in the underlying infrastructure	Process	Vehicle tracking, alerts	Reliability, quality, copyright, coverage, credibility
Mobilitate	Personal, social	Unique ethos, personal enrichment, re-creation, group accomplishment, group attraction, instrumentality, cognitive capital/ self-efficacy, meeting own need, protective, structural capital, relation management, socio-political motives	Structure, process	Service delivery problems	Credibility, social justice
Harassmap	Personal, social	Unique ethos, self-expression, self-image, re-creation, group accomplishment, identity, instrumentality, cognitive capital/ self-efficacy, sense of community, meeting own need, trust in the underlying infrastructure, protective, relation management, socio-political motives	Structure, process	Incidents of sexual harassment	Privacy, credibility, sustainability, social justice

### 8.4.5 Coleman, Georgiadou and Labonte's nature and motivation of producers

Coleman *et al* [2009] considered the nature and motivation of *producers* of volunteered geographical information. They characterized the contributors of VGI as seen by the

## 8. Assessing qualitatively taxonomies of user generated content

early commentators into five overlapping categories:

- *Neophyte*: someone with no formal background in the subject, but with the interest, time and willingness to offer an opinion (or data).
- *Interested amateur*: someone gaining knowledge and expertise in the subject, though reading, experimenting and consulting with other colleagues and experts.
- *Expert amateur*: someone knowing much about a subject and practising it passionately on occasion, but not relying on it for a living. Presumably, this also includes those with detailed and relevant local knowledge about their environment, as opposed to knowledge about a discipline.
- *Expert professional*: someone with the education and professional recognition in the subject to be able to rely on it for a living, and may be sued if they fail their customers.
- *Expert authority*: someone with greater knowledge and experience of the subject than the expert professional, with an established track record and in a position to lose that reputation and even their livelihood if their credibility is lost, even temporarily.

With the members of the ICA's Commission on Geoinformation Infrastructures and Standards, I have used this taxonomy of Coleman *et al* [2009] in a model of the stakeholders in an SDI, in considering the skills of the producers of VGI [Cooper *et al* 2011c]. The ten selected repositories are mapped against these five types of producers in Table 8.6.

Table 8.6: VGI repositories and producers' motivation [Coleman *et al* 2009]

Repository	Neophyte	Interested amateur	Expert amateur	Expert professional	Expert authority
Tracks4Africa		X	X	X	X
OpenStreetMap		X	X	X	X
Wikimapia		X	X	X	X
Google Earth	X	X	X	X	X
Google Maps		X	X		
SABAP2			X	X	X
De Longueville			X	X	X
In-car navigation	X	X	X	X	X
Mobilitate	X	X	X		
Harassmap	X	X	X		

Coleman *et al* [2009] then identified four overlapping contexts in which individuals contribute VGI:

- *Mapping and navigation*: such as Tracks4Africa, OpenStreetMap, Wikimapia, Google Earth, Google Maps and, of course, Vehicle navigation.
- *Social networks*: such as Tracks4Africa, OpenStreetMap, SABAP2 and Harassmap.

## 8. Assessing qualitatively taxonomies of user generated content

- *Civic/governmental*: such as OpenStreetMap (particularly through the Humanitarian OpenStreetMap Team (HOT)), Mobilitate and Harassmap.
- *Emergency reporting*: OpenStreetMap (particularly through HOT), Vehicle navigation, Mobilitate and Harassmap, though it is not certain how quick the responses will be for any of these.

Coleman *et al* [2009] distilled their research to identify the following reasons why producers contribute: *Altruism, Professional or personal interest, Intellectual stimulation, Protection or enhancement of a personal investment, Social reward, Enhanced personal reputation, Outlet for creative and independent self-expression and Pride of place*, and then three negative reasons: *Mischief, Agenda and Malice and/or criminal intent*. However, these are largely independent of the repositories themselves. For example, reasons of *mischief, agenda or malice and/or criminal intent* could be perpetrated against all of the repositories, even De Longueville *et al* [2010b], which is filtered through the research team.

Drawing on the work of several others on Wikipedia, Coleman *et al* [2009] pointed out that contributions could be *Constructive (Legitimate new content, Constructive amendments, Validation and repair or Minor edits and format changes)* or *Damaging (Mass deletes, Nonsense, Spam, Partial deletes, Offensive content or Misinformation)*. Again, these are largely independent of the repositories themselves. However, to varying extents the repositories can deal with the negative reasons for contributing and the damaging contributions.

- As *Tracks4Africa* validates contributions against those from others, it is resistant to an individual attack, but not one by a team.
- *OpenStreetMap* has a suite of tools and editors to check for errors or invalid contributions, but the remedies can only be done *post hoc*.
- *Wikimania* seems to be vulnerable and exploited to a limited extent, but it does also have editors who make corrections *post hoc*.
- *Google Earth* and *Google Maps* have the resources of a very large organisation behind them, so they are likely to be able to take remedial action rapidly. Further, they can probably be very intimidating through legal channels, which probably acts as a deterrent. Nevertheless, Google suspended the use of Google Map Maker in May 2015, because of invalid VGI contributed through it [Kanakarajan 2015; Perez 2015; Siegal 2015].
- For *SABAP2*, while it is easy to submit individual field sheets that are fraudulent, their impact would be minimal. Any significant attack would probably be picked up quickly, though, not only by the *SABAP2* administrators, but also by the contributors to *SABAP2*.
- As all the contributions to De Longueville *et al* [2010b] were filtered and processed by them, they most probably would have detected any significant attack.
- As for *Vehicle navigation* the contributions by each user are limited and hence probably fairly easy to detect and control. Perhaps the only way to cause more than token damage would be to put a lot of the vehicle tracking devices on drones?

## 8. Assessing qualitatively taxonomies of user generated content

---

- All contributions to *Mobilitate* and *Harassmap* were or are followed up with field work, which would detect any significant attacks.

### 8.4.6 Castelein, Grus, Cromptvoets and Bregt's characterization of repositories of VGI

Castelein *et al* [2010] characterized repositories of VGI (though their text implies they characterized VGI *per se*) from the perspective of SDI components, using the conceptual model of Rajabifard *et al* [2002], which has five core components. To this, they added fourteen characteristics to describe VGI, chosen because of ease of measurement by Web survey, objective character and clear presentation of the five SDI components. This taxonomy is good for characterization (their intended purpose) but not for our purposes, as discussed below.

- *Policy*
  1. Whether or not registration is required to contribute.
- *Access network*
  3. If application programming interface(s) are available.
  4. Available services: download and/or upload of data.
- *Standards*
  5. If there are standard feature types and/or standard data formats for uploading.
- *Data*
  6. Total number of contributions uploaded.
  7. Data types that can be uploaded: point, line and/or polygon data.
  8. If the last update or contribution to the Web site within the last hour.
  9. If there is a thematic focus or user community with a specific theme.
  10. Geographic extent of the data: global, continent, region, etc.
  11. If the site only has VGI, or if it is combined with official data.
- *People*
  12. Number of registered users.
  13. Number of unique visitors per day.
  14. Number of unique sites linking to the site.

It is difficult to use numerical characteristics (ie: 6, 12, 13 and 14 above) for classification. Further, there is much overlap between these numerical characteristics, as they all reflect the popularity of the repositories, which is shown to some extent by their Alexa rankings,

## 8. Assessing qualitatively taxonomies of user generated content

see Table 8.1 above. The *standards* component would be much more useful if it included standards for *metadata* and *quality* (see Sections 6.2 and 5.12, respectively), which would still be easy to measure and objective. This taxonomy of Castelein *et al* [2010] is applied to the repositories in Table 8.7. Note that the contribution volume (under **Data** in the table) can be represented in different ways that are not mutually comparable.

Table 8.7: VGI repositories and repository characterization [Castelein *et al* 2010]

Repository	Policy	Access network	Standards	Data	People (Popularity)
<b>Tracks4Africa</b>	Registration required	Upload, limited download	Yes	over 843000 kms of roads, over 135000 listings, over 30000 photos; tracks; current; routes & PoIs; Africa; VGI	Fairly popular
<b>OpenStreetMap</b>	Registration required	API, upload, download	Partially	64GB in XML; all; current; general; world; both	Popular
<b>Wikimapia</b>	Anyone can add; registration required for editing	API, upload, download	No	over 24 million objects; all; current; points of interest; world; both	Fairly popular
<b>Google Earth</b>	Registration required; also scrapes from Wikipedia and Panoramio	API, upload, limited download	Yes (eg: KML)	Many contributions; all; current; general; world; both	Very popular
<b>Google Maps</b>	Registration required	API, limited upload, limited download	Yes	Many contributions; all; current; general; world; both	Very popular

*Continued on next page*

## 8. Assessing qualitatively taxonomies of user generated content

Table 8.7: VGI repositories and repository characterization, *continued*

Repository	Policy	Access network	Standards	Data	People (Popularity)
<b>SABAP2</b>	Registration required	Upload, limited download	Yes	Over 127000 cards and over 571000 incidental sightings; points; current; birds; southern Africa; both	Specialist; nearly 1700 observers
<b>De Longueville</b>	Participants selected	Not applicable	Not applicable	Unknown volume; points & text; project; environment; project; both	Project only
<b>In-car navigation</b>	Registration required	Upload	Yes	Many contributions; points; current; traffic & road conditions; many countries; both	Popular
<b>Mobilitate</b>	Registration required	Upload	Partially	Over 12000 issues reported; points & text; fairly current; service delivery; South Africa; VGI	Limited; 26000 users
<b>Harassmap</b>	Registration not required	Upload	Yes	Over 1400 harassment reports; points & text; current; harassment; Egypt; VGI	Limited

## 8. Assessing qualitatively taxonomies of user generated content

### 8.5 VGI repositories and citizen science

Repositories of VGI can fulfil a variety of purposes, including contributing scientific data. The typology of citizen science of Wiggins & Crowston [2011] is discussed above in Section 4.4.2. They identified five types of citizen science:

1. **Action**, focusing on local issues;
2. **Conservation**, to support stewardship and management of natural resources;
3. **Investigation**, being the classic type of citizen science, collecting data and specimens for scientific research;
4. **Virtual**, using computers and networks entirely, without any physical elements; and
5. **Education**, where education and outreach, both formal and informal, are the primary goals.

In Section 4.4.2, I added to this the following type:

6. **Subject**, where the citizen is the subject of the research, either actively or passively.

As in the sections above, the ten selected repositories of VGI are compared against this extended typology of Wiggins & Crowston [2011] in Table 8.8.

Table 8.8: VGI repositories and citizen science [Wiggins & Crowston 2011]

Repository	Action	Conservation	Investigation	Virtual	Education	Subject
Tracks4Africa			X			
OpenStreetMap			X	X		
Wikimapia			X			
Google Earth			X			
Google Maps			X			
SABAP2	X	X	X			
De Longueville		X	X			
In-car navigation			X			X
Mobilite	X		X			
Harassmap	X		X		X	X

### 8.6 Summary of the qualitative assessment

The five UGC taxonomies consider UGC/VGI as types of data [Gervais 2009], as the products of types of users [Coleman *et al* 2009], as types of repositories [Castelein *et al* 2010], as the results of stimulants [Wunsch-Vincent & Vickery 2007] and as combinations of these [Budhathoki *et al* 2009], and the sixth considers contributions as citizen science [Wiggins & Crowston 2011]. Tables 8.3 to 8.7 illustrate how the five taxonomies of user-

## 8. Assessing qualitatively taxonomies of user generated content

---

generated content and volunteered geographical information identify and differentiate between the ten selected repositories of VGI, and Table 8.8 illustrates how the typology of citizen science does so. The key results are discussed below.

- The taxonomy on drivers of UGC of the Working Party on the Information Economy of the OECD [Wunsch-Vincent & Vickery 2007], see Table 8.3, does differentiate uniquely between all the repositories, though not necessarily according to the real differentiators in the market of the repositories. The taxonomy would be improved by including a driver for the *purpose* of the repository or VGI, which could be the same as the *contribution* of Budhathoki *et al* [2009].
- Gervais's taxonomy for copyright issues [Gervais 2009], see Table 8.4, does not differentiate uniquely between all the repositories, which is not surprising, as it was adding only one dimension (copyright issues) to the OECD taxonomy [Wunsch-Vincent & Vickery 2007], and would differentiate uniquely when combined with the OECD taxonomy. It does show that all the repositories contain VGI, whether *user authored* and/or *user derived*.
- The framework for conceptualizing VGI of Budhathoki, Nedovic-Budic and Bruce [Budhathoki *et al* 2009], see Table 8.5, does differentiate uniquely between all the repositories, even without considering the many types of *motivation*. Unsurprisingly, the most important differentiators are *contribution* and *issues*.
- The taxonomy of the nature and motivation of *producers* of VGI of Coleman, Georgiadou and Labonte [Coleman *et al* 2009], see Table 8.6, does not differentiate uniquely between all the repositories, which is not surprising, as Coleman *et al* [2009] acknowledge that their classes overlap. Nevertheless, it is a useful aspect of VGI to consider, see Cooper *et al* [2011c]; Sinvula *et al* [2013]; Owusu-Banahene *et al* [2013].
- The taxonomy from the perspective of SDI components of Castelein, Grus, Crompvoets and Bregt [Castelein *et al* 2010], see Table 8.7, really only differentiates uniquely between the repositories in terms of their size and popularity. The taxonomy would be improved by adding characteristics to *policy*, such as quality, vetting and pricing, and *standards*, such as those supported.
- The typology of citizen science of Wiggins and Crowston [Wiggins & Crowston 2011], see Table 8.8, does not differentiate uniquely between all the repositories, which is not surprising, as several of them are general repositories. My addition to their typology does improve the differentiation a bit.
- The repository characteristics that I used in Table 8.2 do differentiate somewhat between all the repositories, but this probably should be expected as I selected the repositories and determined the characteristics!

### 8.7 Preliminary taxonomy of user generated content

This is not a synthesis of the published taxonomies discussed above — it was started before they had been read, though it does draw on them. It merely labels the (potential)

## 8. Assessing qualitatively taxonomies of user generated content

classes and as described above in Section 2.4.5.2, labels are inadequate without definitions. However, this is merely an outline and completing it (such as providing definitions) is beyond the scope of this thesis. In all cases, the user generated content could be a combination of these items.

1. **Nature of the contribution:** Text (creative writing, citizen journalism, educational material, thought pieces (blogs)), Image, Audio, Video, Virtual content (eg: in Second Life), Topological data, Spaghetti data, Attribute data, Unknown/Unrecorded
2. **Size of the contribution:** Large, Medium, Small, Unknown/Unrecorded.
3. **Motivation for making the contribution:** Benevolent (contribute to the community, citizen science), Malevolent (trolls, hate speech, snark, vendetta, defamatory, identify theft, misrepresentation/impersonation, phishing, spam, grooming, grief tourism), Advertising, Ego/Vanity/Conceit, Polemic (eg: commentary), Coercion, Solicited (gather data for monitoring service delivery, crowd-seeded), Crowd-sourced, Organising (political events, activism, smart mob, flash mob), Monitoring (fact-checking of politicians, etc), Unknown/ Unrecorded.
4. **Authority for making the contribution** (are there legal definitions for any of these?): Primary source (contributing data about themselves or about what they have witnessed), Secondary source (contributing data obtained from a 'reliable' source (eg: a newspaper article) or documented properly at the time it was obtained from another person), Hearsay (an oral history obtained from others), Legend (folk tale, urban legend, etc), Speculative (assumptions, based on other data, hearsay, etc), Modelled/Simulated, Ignorant source (contributed by someone without the faculties to determine the authenticity of the data, etc), Fake (knowingly false), Unknown/Unrecorded.
5. **Ability to make the contribution:** Competent, Incompetent, Ill-informed, Libelous, Accountable.
6. **Funding for the contribution:** Self-funded/unfunded, Funded by a vested interest, Funded by a benevolent source, Funded by use/users, Unknown/Unrecorded.
7. **Ownership of the content:** Stolen intellectual property, Material out of copyright, Completely original contribution, Redistributed on behalf of someone else, Remixed contribution/mashup, Scraping Web sites, Acknowledgement of sources, Unknown/Unrecorded.
8. **Mechanism used for making the contribution:** Directly onto the public space of a Web site, Through an editor (human/automated, quality of the editing), Pre-production moderation, Post-production moderation, Peer-based moderation (social filtration and accreditation), Gathered by a bot (eg: scraping, a bot that geocodes the data), Unknown/Unrecorded.
9. **Intelligibility of the contribution:** Exemplary legend, Variable, Meaningless squiggles, Unknown/Unrecorded.
10. **Quality of the contribution:** Exemplary, Variable, Self-contradictory, Unknown/Unrecorded.

## 8. Assessing qualitatively taxonomies of user generated content

---

11. **Quality assurance/control of the contribution:** Unverified, Unverifiable (uncheckable, eg: hearsay from a dead person), Peer review (open, blind, etc), Double entry, Vetted by an editor, Unknown/Unrecorded.
12. **Technical quality of the contribution** (beauty is in the eye of the beholder): Exquisite, Low resolution (coarse, murky), High resolution (precise, clear), Literate, Ambiguous, Unknown/Unrecorded
13. **Metadata for the contribution:** Exemplary, Non-existent, Unknown/Unrecorded.
14. **Value of the contribution:** Newsworthy, Value supported by user feedback, Derivative, Valueless, Spam, Well made, Unknown/Unrecorded.
15. **Relevance of the contribution** (SASQAF might be of interest here): Exemplary, Relevance supported by user feedback, Irrelevant, Unknown/Unrecorded.
16. **Ethics of the contribution:** No identified ethical problem, Contravenes the privacy of others, Contravenes legislation (Not necessarily a bad thing!), In poor taste/ Controversial content, Adult content (eg: locations of sex shops — with photos. An ethical dilemma is whether or not it is ethical to consider adult content to be an ethical issue!), Imposes a power relationship (Depends on one's perspective!), Empowers the oppressed (Depends on one's perspective!), Unknown/Unrecorded.
17. **Intellectual property issues:** Well documented, Undocumented, Unknown/ Unrecorded.
18. **Liability for the contribution:** Author accepts liability, Covered by a valid disclaimer, Accompanied by an invalid disclaimer, Unspecified, Unknown/Unrecorded.
19. **Authorship:** Verified expert, Verified amateur with local/domain knowledge, Dubious Anonymous, Unknown/Unrecorded.
20. **Personality of the contribution:** Personal touch, Impersonal, Subjective, Pseudo-objective, Objective, Unknown/Unrecorded.
21. **Currency/validity** (The dating of UGC is of concern): Current version, Deprecated version, Withdrawn version, Out-dated/superseded version, Of historical interest, Disproved version, Unknown/Unrecorded.
22. **The supply chain:** Unknown/Unrecorded.
23. **Number of contributors:** Unknown/Unrecorded.
24. **Diversity of contributors:** Unknown/Unrecorded.
25. **Licensing:** Unknown/Unrecorded.
26. **Maturity:** Unknown/Unrecorded.
27. **How many links in the contribution chain:** Unknown/Unrecorded.
28. **Knowledge of quality/availability of metadata:** Unknown/Unrecorded.
29. **Protocol/rationale for VGI vs free-for-all:** Domain knowledge, GIS knowledge, Moderation/editing as part of the protocol, Unknown/Unrecorded.

---

## 8. Assessing qualitatively taxonomies of user generated content

---

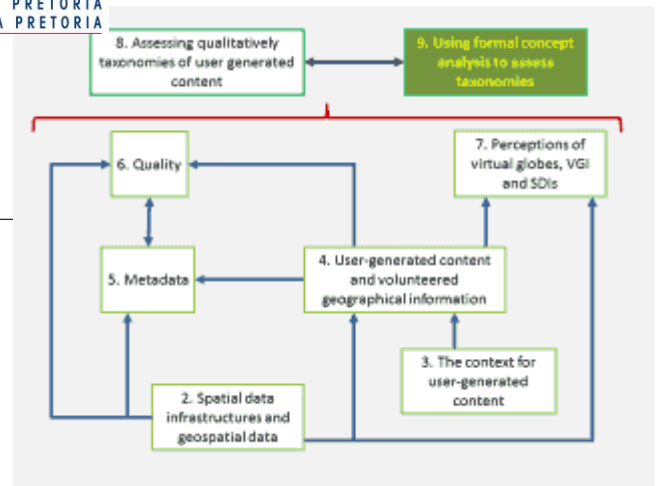
- 30. **Utility:** Unknown/Unrecorded.
- 31. **Quality:** Rigour of the screening based on quality, Availability and type of meta-data, Quality dimensions, Extent to which liability for the data is accepted, Unknown/Unrecorded.
- 32. **Reliability:** Unknown/Unrecorded.
- 33. **Device used to generate and submit the content:** PC, Cell phone, PDA, etc.
- 34. **Medium used to submit the content:** Internet, Wi-Fi, SMS, MMS, etc.

### 8.8 Summary and looking ahead

This chapter has presented a qualitative assessment of various *repositories* of VGI and *taxonomies* of VGI, both separately and against one another. For this, the chapter discussed the five taxonomies of UGC and VGI and one of citizen science which are used for the assessment; the ten repositories containing VGI selected for assessing the taxonomies; the candidate repositories that were not selected; the qualitative assessment itself; and a preliminary taxonomy of user generated content that I developed.

Chapter 9 builds on this chapter and presents a more rigorous analysis of these taxonomies, using *formal concept analysis (FCA)*. It also introduces the concept of *stability exploration* within FCA.

\*\*\*\*



## Chapter 9

# Using formal concept analysis to assess taxonomies

### 9.1 Overview of the chapter

Chapter 8 presented a qualitative assessment of the ability of several taxonomies to discriminate adequately between ten repositories of volunteered geographical information. This chapter takes that further, using *formal concept analysis (FCA)*.

- It presents an overview of *FCA* (Section 9.2), including *stability* and *instability in a lattice* (Sections 9.2.2 and 9.2.3), some *lemmas on stability* in a lattice (Section 9.2.5), *tools* that support *FCA* (Section 9.2.6), and *attribute exploration* (Section 9.2.7).
- It outlines the correspondence between the *feature model* (see Section 2.3.5, Figure 2.2 and Table 2.1) and *FCA*, in Section 9.3.
- In Section 9.4, it explores how *FCA* can be used to assess a taxonomy, covering *discrimination adequacy*, *absent* and *redundant* attributes and objects, and *high intensional* and *extensional stability*. Sections 9.2 and 9.4 formed the basis of a conference paper on *instability*, Cooper *et al* [2010b].
- It introduces *stability exploration* in Section 9.5, including the *rationale* for stability exploration (Section 9.5.2), a *methodology* for implementing it (Section 9.5.3) and some possible *applications* of stability exploration (Section 9.5.4).

## 9. Using formal concept analysis to assess taxonomies

- In Section 9.6, it uses FCA to assess the discrimination adequacy of the taxonomies discussed in Chapter 8. A preliminary version of this analysis, together with the qualitative analysis in Section 8.4, was published as a chapter, Cooper *et al* [2012b], in the book “*Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications*” [Díaz *et al* 2012]. Please note that the quality in general of three of these repositories is assessed in Chapter 6: SABAP2 in Section 6.8.1, OpenStreetMap in Section 6.8.2 and Tracks4Africa in Section 6.8.3.

The major original contributions that I have made that are presented in this chapter are as follows.

- I determined how to use **formal concept analysis** for assessing existing taxonomies, such as to determine their *discrimination adequacy*: previously, FCA has been used to create taxonomies (eg: Kourie & Oosthuizen [1998]), but not to assess them. See Section 9.4.
- In FCA, I determined that there can possibly be value in **instability** in a lattice when assessing a taxonomy [Cooper *et al* 2010b], as the instability could represent extreme values rather than noise, see Sections 9.2.2 and 9.4.
- I contributed a few **lemmas on stability in a lattice**, see Section 9.2.5. Amongst others, these provide lower and upper bounds for intensional and extensional stability indices.
- In FCA, I discovered **stability exploration** and developed a specification of it, see Section 9.5. Stability exploration can possibly be used as a decision support tool, see Section 9.5.4.

Further, the key contributions that I have made that are presented in this chapter are as follows.

- In Section 9.3, I discussed briefly the correlation between FCA and the **feature model**, which had been introduced in Section 2.3.5, Figure 2.2 and Table 2.1.
- I used FCA to **assess** how well several **taxonomies** of user-generated content discriminated between various repositories of volunteered geographical information, see Section 9.6 and Cooper *et al* [2012b].
- I found that the ability to show **sub-contexts** in the FCA tool ConExp [Yevtushenko *et al* 2003], by selecting and deselecting attributes and objects, could be used to find manually more effective combinations of attributes (ie: the classes of the taxonomy being assessed), see Sections 9.6.3 and 9.6.5, and Figures 9.15 and 9.19.

Finally, this chapter raises some questions for further research, see also Section 9.6.8.

- Referring to the lemmas in Section 9.2.5, is there is a stability index value above which every concept is stable, and/or below which every concept is unstable, or does it depend on the application?
- Referring to the methodology for implementing stability exploration in Section 9.5.3, could the stability exploration be stopped automatically before each concept has at

## 9. Using formal concept analysis to assess taxonomies

most one own attribute and one own object, such as when the change in the stability value is below some threshold?

- Referring to the levels of expertise of producers, as identified by Coleman *et al* [2009] (see Section 9.6.4), is it possible to differentiate meaningfully between the contributions of an interested amateur and an expert amateur?

Broadly, the analysis presented in this chapter has two objectives:

1. Illustrate the extent to which there could be value in applying formalisms such as formal concept analysis to taxonomies for geographical information; and
2. Show what is required for a robust taxonomy of user-generated content in general, and of volunteered geographical information in particular: this should aid the users, experts and researchers of UGC with identifying and understanding the content, as well as facilitating other theoretical research on UGC.

## 9.2 Formal concept analysis (FCA)

### 9.2.1 Overview of formal concept analysis

Formal concept analysis (FCA) was invented by Rudolf Wille at the *Technische Hochschule Darmstadt* in Germany, in the early 1980s [Wille 1982; Ganter & Wille 1997; Priss 2006]. Essentially, it uses a lattice of formal *concepts* with *objects* and *attributes*, and the linkages between them, for data analysis, knowledge representation and information management [Priss 2006]. The following overview of FCA has been derived from several sources: Wille [1982]; Ganter & Wille [1997]; Burmeister [2003]; Yevtushenko *et al* [2003]; Carpineto & Romano [2004]; Priss [2006]; Ganter [2007]; Kuznetsov [2007]; Bělohlávek [2008]; Klimushkin *et al* [2010] and Wikimedia [2016].

A *lattice* is a *partially ordered set* (poset), denoted as  $(P; \leq)$ , where for any pair of elements  $x$  and  $y$  in  $P$ , both the *supremum* (the least upper bound, also known as the *join*), denoted by  $x \vee y$ , and the *infimum* (the greatest lower bound, also known as the *meet*), denoted by  $x \wedge y$ , always exist. Such an ordered set is a *complete lattice* if the supremum, denoted by  $\vee S$ , and the infimum, denoted by  $\wedge S$ , exist for any subset  $S$  of  $P$ . A complete lattice always has a top element, an element that is greater than all the other elements (known as the *unit*), and the dual of this, namely a bottom element, an element that is smaller than all the other elements (known as the *zero*). The *unit* is the *supremum* and the *zero* the *infimum* for the entire lattice. In an ordered set, element  $x$  is *covered* by  $y$  if  $x < y$  and there is no  $z \in P$  such that  $x < z < y$ . This is denoted as  $x < y$ . The inverse of this is that  $y$  *covers*  $x$  and it is denoted as  $y > x$ .

Every finite poset  $(P; \leq)$  can be drawn and if  $x < y$  and if element  $x$  is placed below element  $y$ , then the diagram is known as a *line diagram* or a *Hasse diagram*. Figure 9.1 shows a line diagram of a complete lattice, where node 1 is the *zero* of the lattice, and node 7 the *unit*. Node 5 covers node 2 and node 2 is covered by node 5, for example. As this

## 9. Using formal concept analysis to assess taxonomies

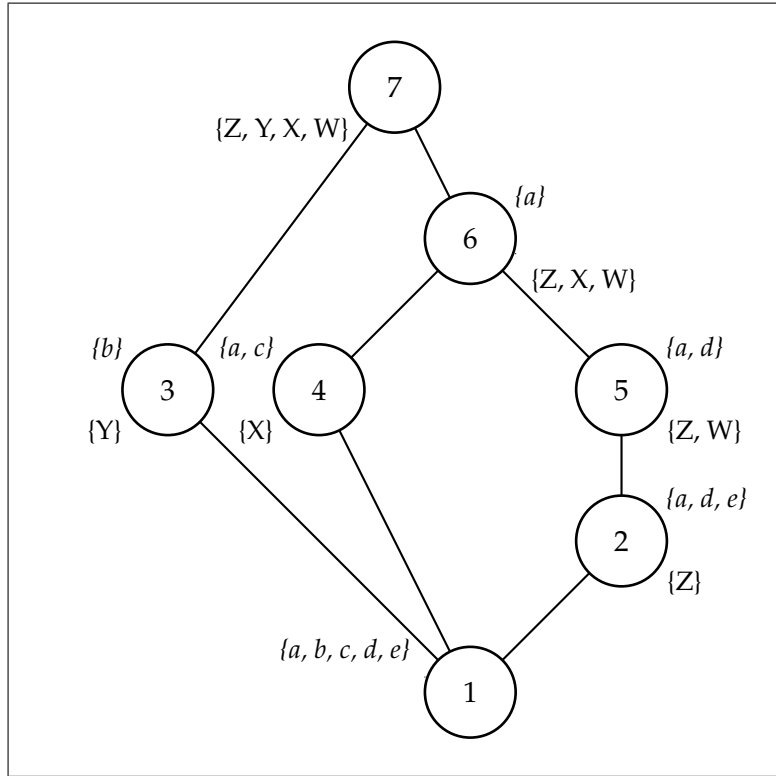


Figure 9.1: An example of a line diagram

is a *formal concept lattice*, attached to each node are *objects* (the upper-case letters) and *attributes* (the lower-case letters), as discussed below. Please note that any one lattice could be drawn in many different ways: there is no perfect way to produce a line diagram. This is illustrated by Figure 9.2, which shows the same lattice as in Figure 9.1, but drawn differently.

In FCA, the notation for a *formal context* is:  $\mathbb{K} := (G, M, I)$ , where  $I$  is the binary relation between the sets of objects,  $G$ , and attributes,  $M$ , namely:  $I \subseteq (G \times M)$ . Please note that the  $G$  comes from the German for object, *Gegenstände*, and the  $M$  from the German for attribute, *Merkmale*. Further, a *concept* in FCA is termed *formal* as it aligns with the classical theory of concepts, as opposed to non-classical theories of concepts, or non-mathematical variants of the classical theory (eg: in psychology or philosophy).

Each node (or element) of the formal concept lattice corresponds to a pair  $(A, B)$ , where  $A \subseteq G$  and  $B \subseteq M$ . The *derivative* of the set of objects,  $A$ , is denoted by  $A'$  and is the set of *all* attributes that are common to all objects in  $A$ . The derivative of the set of attributes,  $B$ , is dually defined and denoted by  $B'$ .

If  $A \subseteq G$  and  $B \subseteq M$ , the *Galois connection* is given by the following *derivation operators*:

$$A' := \{m \in M \mid gIm \ \forall \ g \in A\} \quad (9.1)$$

$$B' := \{g \in G \mid gIm \ \forall \ m \in B\} \quad (9.2)$$

## 9. Using formal concept analysis to assess taxonomies

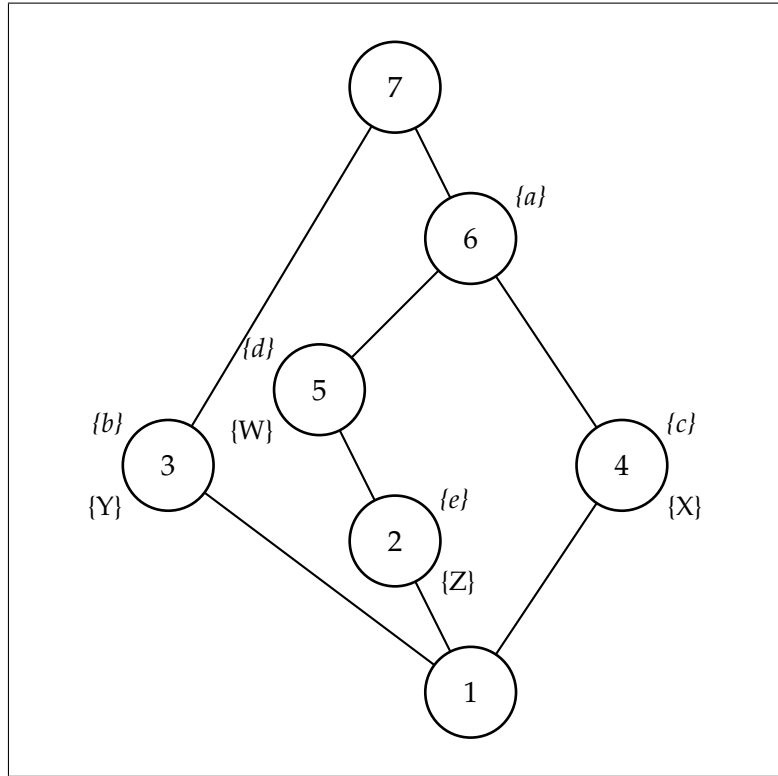


Figure 9.2: The lattice shown in Figure 9.1, but with reduced labelling and a different layout

$\mathfrak{C} = (A, B)$  is a formal *concept* in  $\mathbb{K}$  if, and only if,  $B$  contains all the attributes that the objects in  $A$  have in common, and only those attributes, that is, if  $A' = B$ , and  $A$  also contains all the objects that share the attributes in  $B$ , and only those objects, that is, if  $B' = A$ . For a concept  $\mathfrak{C} = (A, B)$ ,  $B$  is known as the *extent* of the concept (written as  $\text{Ext}(\mathfrak{C})$ ) and  $A$  is known as the *intent* of the concept (written as  $\text{Int}(\mathfrak{C})$ ). If  $\mathfrak{C}_1 = (A_1, B_1)$  and  $\mathfrak{C}_2 = (A_2, B_2)$  are concepts of  $\mathbb{K}$ , then  $\mathfrak{C}_1 \leq \mathfrak{C}_2$  iff  $A_1 \subseteq A_2$  and  $B_2 \subseteq B_1$ <sup>1</sup>. This is known as the ordering on *concepts*.

A concept's *own objects* are those in its extent that are not in the extent of any *sub-concept*, and a concept's *own attributes* are those in its intent that are not in the intent of any *super-concept*. We use the notation  $\xi_e(\mathfrak{C})$  for a concept's set of own objects and  $\xi_i(\mathfrak{C})$  for a concept's set of own attributes.

It can become messy in a line diagram to label each node with all the objects and attributes in its derivatives, so *reduced labelling* is used, as is shown in Figure 9.2. Each object and each attribute is entered only once in the diagram, as an *own object* or *own attribute*, that is, at its “first occurrence” — the lowest concept for an object, and the highest for an attribute. Thus, for example, in Figure 9.2, concept 2 has the extent of  $\{W, Z\}$  and the intent of  $\{a, e\}$ , that is, it is the concept  $(\{W, Z\}, \{a, e\})$ . In a line diagram, the top half of the circle for the node could be filled in if that concept has an own attribute and the

<sup>1</sup>Please note that here,  $A_1 \subseteq A_2$  is equivalent to  $B_2 \subseteq B_1$ .

## 9. Using formal concept analysis to assess taxonomies

bottom half if it has an own object, as shown in Figures 9.4 and 9.7, for example.

A formal *context* can also be considered as a table, relating objects to attributes — indeed, FCA tools such as Concept Explorer (ConExp) [Yevtushenko *et al* 2003] and Con-Imp (from Contexts and Implications) [Burmeister 2003] use a table for input to derive the concepts of the lattice and produce associated line diagrams according to selected drawing criteria. For example, Table 9.1 shows the formal context in Figures 9.1 and 9.2, but as a cross-table — a cross in cell  $ij$  indicates that object  $i$  (in row  $i$ ) is described by attribute  $j$  (in column  $j$ )).

Table 9.1: A cross-table of the formal context shown in Figures 9.1 and 9.2.

	a	b	c	d	e
Z	×			×	×
Y		×			
X	×		×		
W	×			×	

The notation for a *formal context* is given above as:  $\mathbb{K} := (G, M, I)$ . Strictly speaking, though, this is a *one-valued context*, in that each attribute has one of only two values, *present* or *absent*. A formal context can also be a *many-valued context*, that is, have attributes that can have multiple values (eg: a date or a count). A *many-valued context* is a quadruple,  $\mathbb{K} := (G, M, V, I)$ , where  $G$  is a set of objects,  $M$  a set of many-valued attributes,  $V$  a set of attribute values and  $I$  the ternary relation,  $I \subseteq (G \times M \times V)$ , such that

$$(g, m, v) \in I \text{ and } (g, m, w) \in I \implies v = w \quad (9.3)$$

The expression  $(g, m, v) \in I$  is read as the attribute  $m$  has the value  $v$  for object  $g$ , which could also be written as  $m(g) = v$ . Any many-valued context can be transformed to a one-valued context by replacing every valid attribute and attribute value pair by a new attribute, as is shown in Table 9.2. This process is known as *conceptual scaling*. For the analysis here in Section 9.6, one-valued contexts have been used, not just because it is what ConExp supports, but also because the lattices are clearer, even though they have more attributes.

Table 9.2: A many-valued context and the equivalent one-valued context.

Many	year	number
First	2010	1
Second	2008	2
Third	2008	1
Fourth	2009	2

One	year 2008	year 2009	year 2010	1	2
First			×	×	
Second	×				×
Third	×			×	
Fourth		×			×

## 9. Using formal concept analysis to assess taxonomies

Priss [2006] suggests that FCA's implementation of the notions "extension" and "intension" differs slightly from their use for a concept in philosophy, such as the "*Sinn*" (*sense*) and "*Bedeutung*" (*reference*) of Frege [1892]. In Frege's example, *morning star* and *evening star* are different *senses* (in the intention of the concept) for the same *referent* (in the extension of the concept), namely *Venus*. Priss [2006] states that "in FCA, an extension occurring with respect to one formal context can have only one corresponding intension, not two different intensions", with the resolution being either to prevent *morning star* and *evening star* from being in the intent of the same formal context, or defining the intent as the set of *morning star* and *evening star*. I would suggest that the latter is a many-valued formal concept that could be transformed into a one-valued context using conceptual scaling.

### 9.2.2 Stability and instability in a lattice

The *stability* of a formal concept is an indication of how much its intent depends on individual objects in the extent, and/or how much the extent depends on individual attributes in the intent [Kuznetsov 2007]. Thus, the *intensional stability* of a concept indicates how much its intent depends on individual objects in its extent. It is a measure of the likelihood that removing a random set of objects from the concept's extent would change its intent. In other words, the higher the intensional stability of a concept, the greater the likelihood that any set of one or more attributes is shared by more than one object, and hence is less dependent for its continued existence on any one of those objects remaining in the context.

Similarly, *extensional stability* indicates how much its extent depends on individual attributes in its intent. Again, it is a measure of the likelihood that removing a random set of attributes from the concept's intent would change its extent. A concept with a high stability index exhibits *stability* (ie: is stable), while a concept with a low stability index exhibits *instability* (ie: is instable<sup>2</sup>). Formally, the *intensional stability index*,  $\sigma_i$ , and the *extensional stability index*,  $\sigma_e$ , of concept  $(A, B)$ , are defined in Klimushkin *et al* [2010] as follows:

$$\sigma_i(A, B) = \frac{|\{C \subseteq A \mid C' = B\}|}{2^{|A|}} \quad (9.4)$$

$$\sigma_e(A, B) = \frac{|\{D \subseteq B \mid D' = A\}|}{2^{|B|}} \quad (9.5)$$

Each concept  $\mathfrak{C} = (A, B)$  has  $|A|$  objects in its extent and hence there are  $2^{|A|}$  subsets of  $A$  in its extent. The intensional stability index is the proportion of the  $2^{|A|}$  subsets of  $A$  that have the following property: the attributes  $C'$  shared by the objects in any such subset, say  $C$ , correspond to the concept's intent, that is,  $C' = B$ .

<sup>2</sup>The use of *instable* as opposed to *unstable* is now rare in English [Oxford English Dictionary Department 1973], but it makes sense to use it here to correlate better with *instability*.

## 9. Using formal concept analysis to assess taxonomies

The intent of concept  $(A, B)$  may therefore be characterised as stable if the lattice is characterised by the fact that at least one subset of  $A$ , say  $C$ , is such that a new lattice has the concept  $(C, B)$  instead of  $(A, B)$  if it is constructed from the context  $\mathbb{K}_C := ((G \setminus A) \cup C, M, I_C)$  where  $\forall (x, y) \in I_C : (x, y) \in I$ . The more such subsets of  $A$  can be found, the more stable is the intent of  $(A, B)$  and the closer  $\sigma_i(A, B)$  is to 1. Conversely, if no proper subset of  $A$  has this property, then the concept  $(A, B)$  could be said to be “intentionally” unstable in that  $\sigma_i(A, B)$  is at its minimum possible value.

The notion of extensional stability is similar, but with the roles of extents and intents reversed: the extent of concept  $(A, B)$  is stable with respect to an intent subset,  $D$ , if  $A$  is retained as the extent of a concept in a revised lattice whose attributes are  $D$  rather than  $B$ . The number of such subsets,  $D$ , relative to the total number of possible subsets of  $B$  provides a stability measure for the concept.

### 9.2.3 Examples of stability

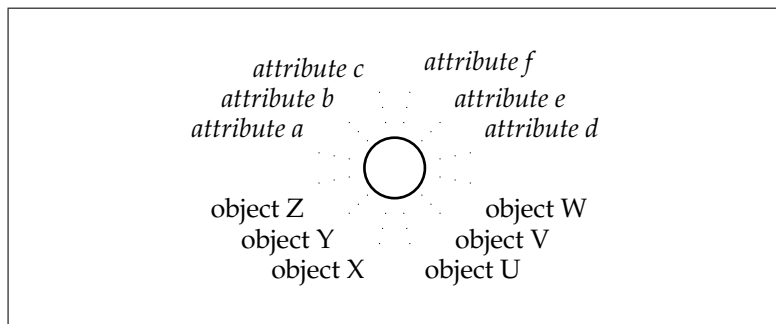


Figure 9.3: A very stable, but rather boring, lattice.

Unsurprisingly, the more objects (or attributes) covered by a formal concept, the more likely it will be intensionally (or extensionally) stable, because of the greater likelihood of “redundant” objects (or attributes). Figure 9.3 shows a lattice that is very stable (with all the objects and attributes being own objects and own attributes of the single concept in the lattice), because a lattice built from any subset of attributes (objects) would yield a concept whose extent (intent, respectively) is unchanged from that of the concept in Figure 9.3. That is,  $\sigma_i(A, B) = 1$  and  $\sigma_e(A, B) = 1$ . Please note that while the empty set is a subset of both the extent and the intent of the concept in Figure 9.3, the concept is both the zero and the unit of the context, so a sub-concept with only the empty set in its extent still has the same intent as the concept, and vice versa. See Section 9.2.5 for a presentation of several lemmas on stability.

However, while Figure 9.3 shows an extreme example of stability, it does illustrate why stability can mean redundancy amongst objects and/or attributes, and why stability can be considered “boring” in some applications because of the low information content<sup>3</sup>. I

<sup>3</sup>In comparison, Kourie & Oosthuizen [1998] use the label “boring” for *concepts* in the lattice that identify associations between attributes that are well known and expected. These concepts can be pruned out to improve machine learning in some cases.

## 9. Using formal concept analysis to assess taxonomies

appreciate that when FCA is being used for applications such as machine learning (eg: Kourie & Oosthuizen [1998]), if concepts in the resulting lattice have high stability it means that the input data were robust with little noise (eg: caused by coding errors or instrument accuracy). Instability (ie: low stability) represents noise that clouds such analysis of the data.

Here, rather than using FCA to classify data (as is done in machine learning, for example), I use FCA to assess taxonomies, with the classes in a taxonomy being attributes in the concept lattice and the things classified by the taxonomy (repositories, in this case) being objects in the concept lattice. Specifically, I assess the ability of the taxonomies to discriminate between the repositories, that is, their *discrimination adequacy*, see Section 9.4.1. For example, if there is high extensional stability then some of the taxonomy's classes are essentially indistinguishable from one another.

While a taxonomy can be created automatically to sift out information from noise in massive data sets obtained from sensors (ie: for machine learning), in practice, taxonomies used by people need to cater for many subtleties, as discussed above in Section 2.4.1. Hence, the taxonomies assessed here and in Chapter 8 could not have been created automatically, such as by using FCA.

### 9.2.4 Intents and extents

Please note that not every subset of an extent forms an extent for some other concept. For example, in Figure 9.4, there is no concept whose extent is  $\{Alice, Bill\}$ , because the derivative of  $\{Alice, Bill\}$  is  $\{Tanned\}$ , but the concept whose intent is  $\{Tanned\}$  has as its extent  $\{Alice, Bill, Amy\}$ .

Further, the derivative of a subset of a concept's extent might be different from the concept's intent, because it might include attributes that are not in the intent of the concept. For example, referring to the concept  $(\{Alice, Bob, Amy\}, \{Tall, Young\})$  in Figure 9.4, the subset of the extent  $\{Amy\}$  has as its derivative  $\{Tall, Tanned, Young, Lovely, Big\}$ , which differs from the intent of the concept in question. The intentional stability of the concept  $(\{Alice, Amy\}, \{Tall, Tanned, Young, Lovely\})$  (which we shall term  $(A, B)$  here) is 0.25. There are 4 subsets in the extent, with their derivatives as shown in Table 9.3. Note that there is only one subset of  $\{Alice, Amy\}$  that has the same derivative as the concept  $(A, B)$ , namely  $\{Alice, Amy\}$  itself. This is because the derivative of  $\{Alice\}$  also includes the attribute  $\{Bright\}$ , the derivative of  $\{Amy\}$  also includes  $\{Big\}$  and the derivative of  $\emptyset$  is the extent of the zero, that is, all the attributes in the context:  $\{Tall, Tanned, Young, Lovely, Bright, Big, Fat\}$ .

### 9.2.5 Lemmas on stability in a lattice

Given the definitions of the intensional and extensional stability indexes in Klimushkin *et al* [2010], shown above in Equations 9.4 and 9.5, I propose the following lemmas on stability in a lattice. Please note that transposing the objects and attributes might seem counter-intuitive, but it is permissible in FCA because the analysis is on the structure

## 9. Using formal concept analysis to assess taxonomies

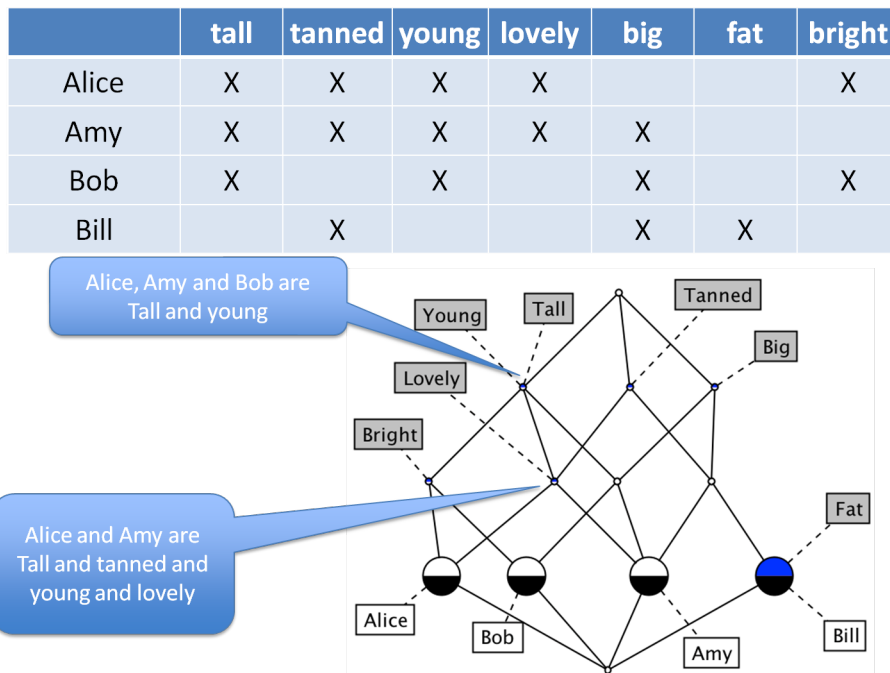


Figure 9.4: The girls of Ipanema [Watson *et al* 2012], adapted from [Jobim *et al* 1962/1964].

Table 9.3: The subsets of the concept covering  $\{Alice, Amy\}$ , with their derivatives.

Subset	Derivative	Equal to the intent?
$\{Alice, Amy\}$	$\{Tall, Tanned, Young, Lovely\}$	Yes
$\{Alice\}$	$\{Bright, Tall, Tanned, Young, Lovely\}$	No
$\{Amy\}$	$\{Tall, Tanned, Young, Lovely, Big\}$	No
$\emptyset$	$\{Tall, Tanned, Young, Lovely, Bright, Big, Fat\}$	No

of the lattice, not the semantics, and one can then use existing tools in new ways, as discussed in Section 9.2.7.

**Lemma 1:** *The intensional stability index of a concept is equal to the extensional stability index of the dual concept when the context has been transposed. This can be written as  $\sigma_i(A, B) = \sigma_e(B, A)$ .*

**Remark 1:** This will also apply to the intensional and extensional stability *values*, that is, the numerators in Equations 9.4 and 9.5, respectively. As a consequence of this, by transposing a context the techniques and tools for conducting *attribute exploration* can be used for *object exploration*, as discussed in Section 9.2.7.

## 9. Using formal concept analysis to assess taxonomies

**Lemma 2:** *The intensional stability of the zero of a lattice is always 1.0.*

**Remark 2:** The zero of a lattice is the bottom-most concept, the concept that is smaller than all the other concepts. The result is that the intent of the zero consists of all the attributes in the lattice. Since the attribute set of every object in the extent of the zero (say,  $A$ ) will be the intent of the zero (say,  $B$ ), therefore  $\sigma_i(A, B) = 1.0$ .

**Lemma 3:** *The extensional stability of the unit of a lattice is always 1.0.*

**Remark 3:** The unit of a lattice is the top-most concept, the concept that is greater than all the other concepts. As this is the dual of Lemma 2, and given Lemma 1, the same argument holds, and so  $\sigma_e(A, B) = 1.0$ .

**Lemma 4:** *The lowest possible intensional stability index for a concept in a lattice is the number of subsets of its own objects divided by the total number of sets of objects in its extent, that is:*

$$\sigma_i(A, B) \geq \frac{2^{|\zeta_i(A)|}}{2^{|A|}} \quad (9.6)$$

**Remark 4:** Each subset of a concept's own objects also has the same derivative as that concept's intent. For example, in Figure 9.5, consider node 9, the concept  $(\{Obj1, Obj2, Obj10\}, \{Attr1, Attr2, Attr4, Attr5, Attr7, Attr10\})$ : the three subsets  $\{Obj1, Obj10\}$ ,  $\{Obj2\}$  and  $\{Obj10\}$  are all own objects of the concept and all have derivatives that are the same as the concept's intent, namely  $\{Attr1, Attr2, Attr4, Attr5, Attr7, Attr10\}$ .

**Lemma 5:** *The lowest possible extensional stability index for a concept in a lattice is the number of sets of its own attributes divided by the total number of sets of attributes in its intent, that is:*

$$\sigma_e(A, B) \geq \frac{2^{|\zeta_e(B)|}}{2^{|B|}} \quad (9.7)$$

**Remark 5:** As this is the dual of Lemma 4, and given Lemma 1, the same argument holds. For example, referring to Figure 9.5 and the concept  $(\{Obj1, Obj2, Obj10\}, \{Attr1, Attr2, Attr4, Attr5, Attr7, Attr10\})$  again: the three subsets  $\{Attr4, Attr5\}$ ,  $\{Attr2, Attr5\}$  and  $\{Attr7\}$  are all own attributes of the concept and all have derivatives that are the same as the concept's extent, namely  $\{Obj1, Obj2, Obj10\}$ .

Together, Lemmas 4 and 5 can be summarized as follows: for any concept  $(A, B)$ , since  $0 \leq |\zeta_i(A)|$  and since  $0 \leq |\zeta_e(B)|$ , therefore:

$$\frac{1}{2^{|A|}} \leq \frac{2^{|\zeta_i(A)|}}{2^{|A|}} \leq \sigma_i(A, B) \leq 1 \quad (9.8)$$

$$\frac{1}{2^{|B|}} \leq \frac{2^{|\zeta_e(B)|}}{2^{|B|}} \leq \sigma_e(A, B) \leq 1 \quad (9.9)$$

## 9. Using formal concept analysis to assess taxonomies

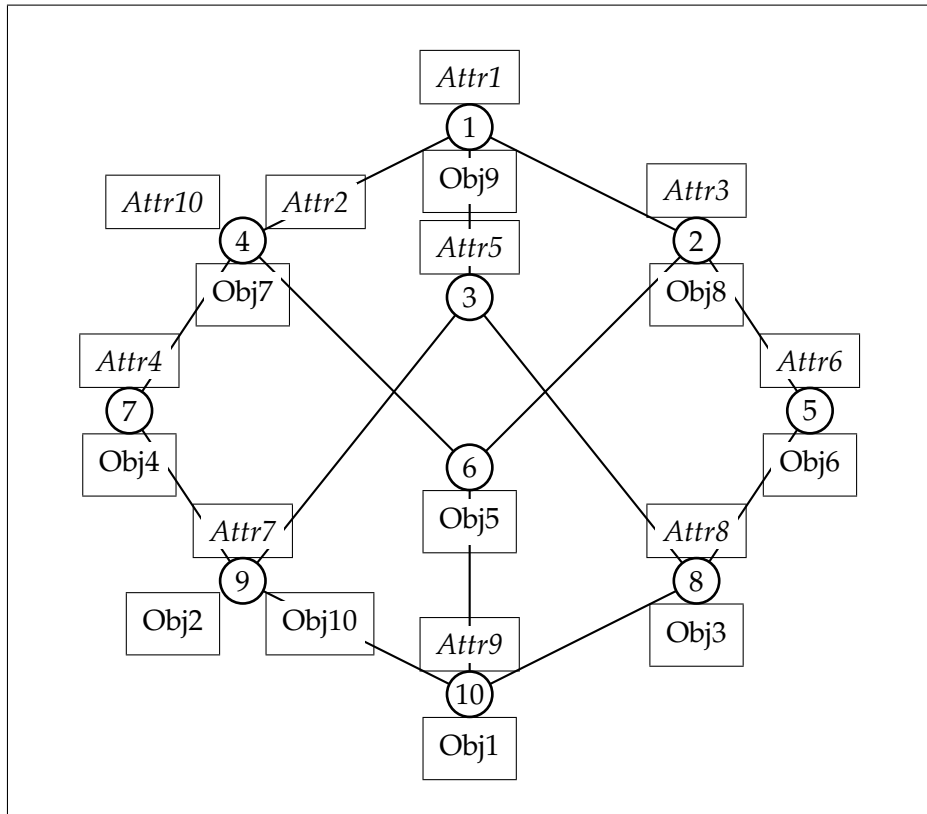


Figure 9.5: A lattice used to show lower bounds for stability indices.

**Lemma 6:** *With the exception of the unit of a lattice, the extensional stability of any concept in a lattice is always less than 1.0.*

**Remark 6:** If concept  $C_1$  is not the unit, then there is at least one other concept,  $C_2$ , that is also not the unit and  $C_2$  has at least one object in its extent that is not in the extent of  $C_1$ , so the extensional stability of  $C_1$  is less than 1.0.

**Lemma 7:** *With the exception of the zero of a lattice, the intensional stability of any concept in a lattice is always less than 1.0. This is the dual of Lemma 6.*

**Remark 7:** As this is the dual of Lemma 6, and given Lemma 1, the same argument holds.

Together, Lemmas 4 and 7 provide the lower and upper bounds for any intensional stability index, and Lemmas 5 and 6 provide the lower and upper bounds for any extensional stability index. This raises the question of whether or not there is a stability index value above which every concept is stable, and/or below which every concept is unstable, or whether or not it depends on the application.

## 9. Using formal concept analysis to assess taxonomies

---

### 9.2.6 Tools supporting formal concept analysis

The tool used to support FCA for this analysis is Concept Explorer (ConExp) [Yevtushenko *et al* 2003]<sup>4</sup>. It was selected because it is an open-source tool, is robust and is used by other researchers in the Department of Computer Science at the University of Pretoria. See, for example, Cleophas *et al* [2006]; Roth *et al* [2008a]; Obiedkov *et al* [2009]; Chan [2010a]), and hence has a pool of expertise that is readily available to us. ConExp provides functionality such as the following.

- *Context editing.*  
Typically a matrix, similar to the illustration in Table 9.1, is used to assemble and edit the lattice quickly. ConExp also allows one to fill, clear or invert a selected block of cells.
- *Building concept lattices from context.*  
Once the matrix has been completed, the tool builds the lattice and the line diagram, which can be adjusted manually to improve its legibility. One can build the lattice for any sub-context by selecting and deselecting attributes and objects. There are also several different layout algorithms that can be applied to improve the look of the line diagram. The labelling of the nodes can be adjusted and the nodes and edges can be drawn to reflect factors such as the size of the extent (ie: number of objects in the extent) or the stability. It is also possible to highlight selections of the lattice, such as the *filter* of a concept (all nodes reachable by ascending paths to the top of the lattice, ie: the intent), or the *ideal* of a concept (all nodes reachable by descending paths to the bottom of the lattice, ie: the extent).
- *Clarification and reduction.*  
This removes redundant attributes and/or objects from the context.
- *Finding the so-called Duquenne-Guigues base of implications that are true in context.*  
Such a base has a minimal possible number of implications amongst all possible bases of implications, that hold in context. Each identified *implication* holds for a set of objects and has a *premise* (a list of zero or more attributes) and a *conclusion* (a list of one or more attributes), such that if all the attributes in the *premise* occur, then so do all the attributes in the *conclusion*. Each *implication* also shows for how many objects the implication holds and whether or not there are objects in context that supports the implication.
- *Finding bases of association rules that are true in context.*  
Effectively, *associations* are *implications* with *non-strict* (or *association*) rules added, which are where only a percentage of the *conclusion* is implied by the *premise*. The approximate rules are also known as the Luxenburger base.
- *Performing attribute exploration.*  
This is an interactive process to see if each *implication* (the “linked” *premise* and *conclusion*) for a set of objects can also apply to other objects, that is, to objects not in the context of the implication. A question is asked of the user about a dependency

---

<sup>4</sup>Note: ConExp’s author requests that users cite his Russian text, Yevtushenko [2000].

## 9. Using formal concept analysis to assess taxonomies

between different attributes from some fixed set of attributes (ie: the *exploration*). If a dependency does not hold, then the user is asked to provide a counter-example [Yevtushenko *et al* 2003] — effectively, the user must add a new object. The next question is asked of the user and this is repeated until the user stops the exploration, or has answered yes to all questions and hence obtained the set of all implications that describe the dependencies between the different attributes in the domain of interest [Yevtushenko *et al* 2003]. Attribute exploration can then reveal “missing” attributes, “missing” objects, “redundant” attributes and “redundant” objects. As attribute exploration is the key functionality provided by ConExp for our analysis, it is discussed in more detail with an example, in Section 9.2.7 below.

### 9.2.7 Attribute exploration

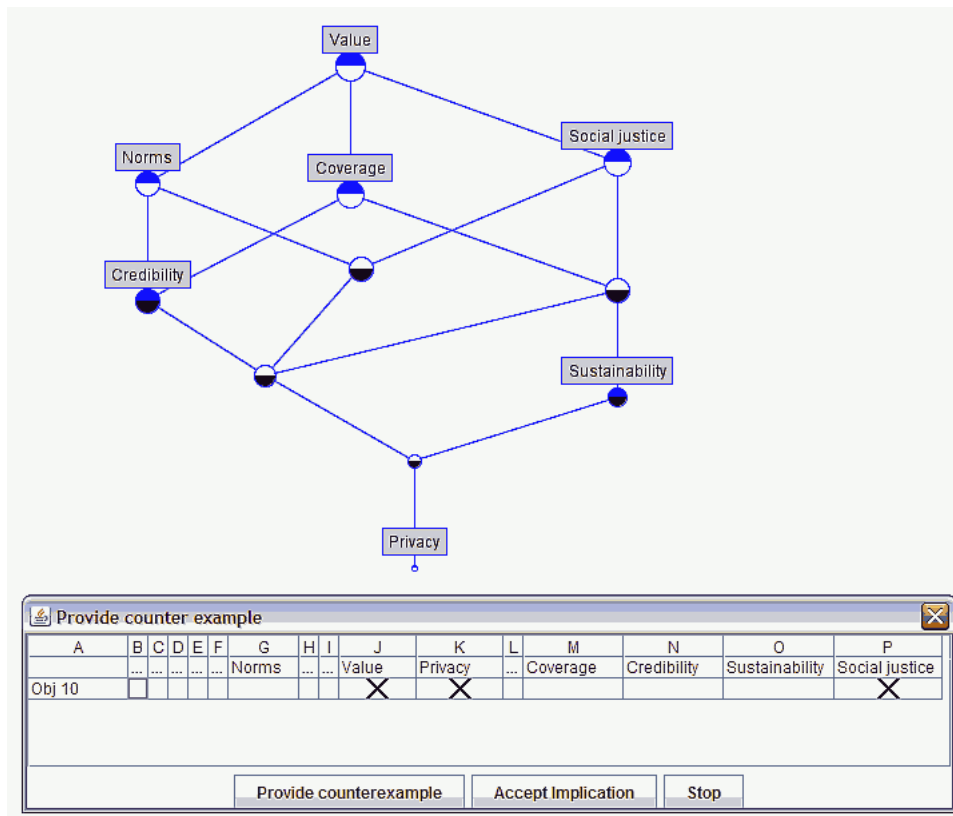


Figure 9.6: Attribute exploration in ConExp [Yevtushenko *et al* 2003].

Figure 9.6 shows a screen shot of ConExp doing attribute exploration. This shows a subset of seven classes of *Issues* from the taxonomy of Budhathoki *et al* [2010]. While the objects (repositories) are not shown explicitly in Figure 9.6, the nodes are scaled to show how many objects each has in its extent. The zero is shown as a point because it has no objects in its extent, and the unit is shown as the largest circle, as it has all the objects in its extent. Please note also that the actual repositories were omitted from the

## 9. Using formal concept analysis to assess taxonomies

---

figure for clarity and because this was not meant to have been a definitive analysis of these repositories at that stage and the repositories had not been classified in detail. The purpose of Figure 9.6 is to illustrate how FCA can provide insights when assessing a taxonomy.

Figure 9.6 illustrates the key breakthrough I made in using FCA for analysing taxonomies, which was revealed to me through attribute exploration. It highlights that *Privacy* did not classify any of the repositories then being considered in the analysis. The zero of the lattice has *high extensional stability* as it has an own attribute but no objects in its extent, and one can see that the attribute *Privacy* is attached to the zero of the lattice. Since the extent of the zero in this lattice is  $\emptyset$ , it has no objects in its extent and hence none of the repositories used in the analysis have *Privacy* as an attribute. In retrospect, preserving privacy is well known as a major concern with UGC/VGI repositories, which confirms the result of the FCA [Cooper *et al* 2010b]. See also Section 9.6.3 for further discussion on this lattice and the taxonomy of Budhathoki *et al* [2010].

In doing this attribute exploration shown in Figure 9.6, the user was asked to provide a counter example of an object with *Privacy* and *Value* as attributes, with the default name of *Obj10* for the object. The user was providing the counter example (though it had not yet been named), while specifying an additional attribute, *Social Justice*, for the object.

In terms of the *duality principle*, the objects and the attributes can be swapped in such a tool for analysis (ie: the axes of the matrix transposed), if such analysis would be useful to explore. In practice, this would put the objects (ie: the repositories in our case) into the intent of the concept and the attributes (ie: the taxonomy classes) into the extent. Effectively, this would allow one to do *object exploration* using the *attribute exploration* function of Conexp. The object exploration would then allow one to determine the similarity of different objects, for example.

The implications of attribute exploration in the context of assessing taxonomies is discussed below in Sections 9.4 and 9.6. Section 9.5 draws on the technique of attribute exploration for *stability exploration*.

Obviously, there is much more to FCA than this outline given above in Section 9.2, but those additional theorems and constructs are not relevant here.

### 9.3 FCA and the feature model

As mentioned above in Section 2.3.5, Derrick G Kourie [2014, pers comm] has proposed that there is a correspondence between FCA and the feature model used for representing geospatial data. There are two ways to approach this, which are shown in Figures 9.7 and 9.8.

1. At the *abstract* level, each *feature* is an object in an FCA lattice. Each attribute in the lattice is then any one of the other concepts shown in Figure 2.2 and Table 2.1, namely *feature concept*, *feature type*, *feature instance*, *non-spatial attribute*, *attribute value*

## 9. Using formal concept analysis to assess taxonomies

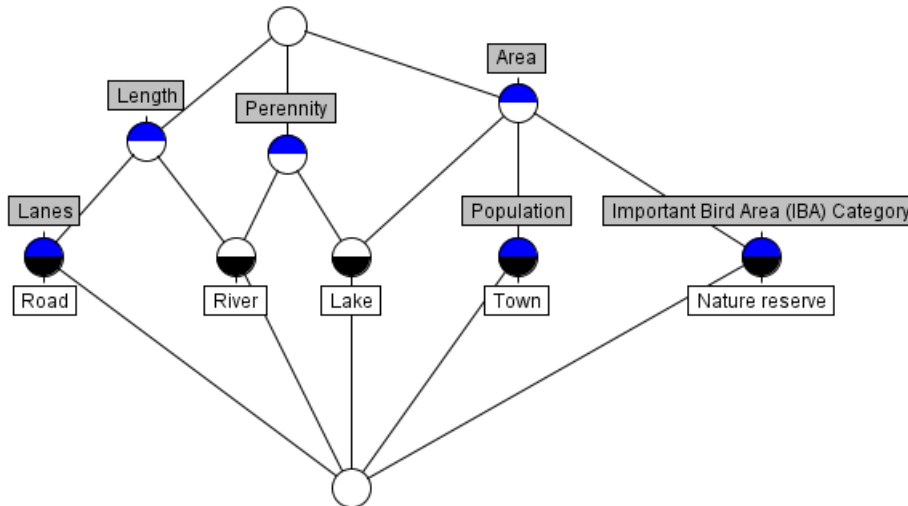


Figure 9.7: Feature model in FCA

*domain, attribute value, association, operation, spatial attribute, geometry, topology, alternate spatial attribute, symbology or metadata.* This is illustrated in Figure 9.7.

2. At the *instance level*, each *feature instance* or *exemplar* is an object in an FCA lattice, as Kourie proposed. Each attribute in the lattice is then an instance of any one of the other concepts shown in Figure 2.2 and Table 2.1, namely *feature concept, feature type, non-spatial attribute, attribute value domain, attribute value, association, operation, spatial attribute, geometry, topology, alternate spatial attribute, symbology or metadata.* This is illustrated in Figure 9.8.

### 9.4 Applying formal concept analysis to assess taxonomies

“Formal Concept Analysis (FCA) is a method for data analysis, knowledge representation and information management” [Priss 2006] that can also be used in information retrieval and machine learning. For example, a lattice can be used to structure a search space into clusters of objects (such as documents), which could use the clusters for improving inadequate queries. This clustered search space could also integrate queries and browsing: find a node, move to a related node and then maybe prune the lattice by refining the query. FCA could also be used to reorganise the results of a general query on a massive search, such as from an Internet-wide search engine [Priss 2006].

However, I have not used FCA to analyse, represent or classify data, but to assess the *discrimination adequacy* of taxonomies for *user-generated content* in general, or for *volunteered*

## 9. Using formal concept analysis to assess taxonomies

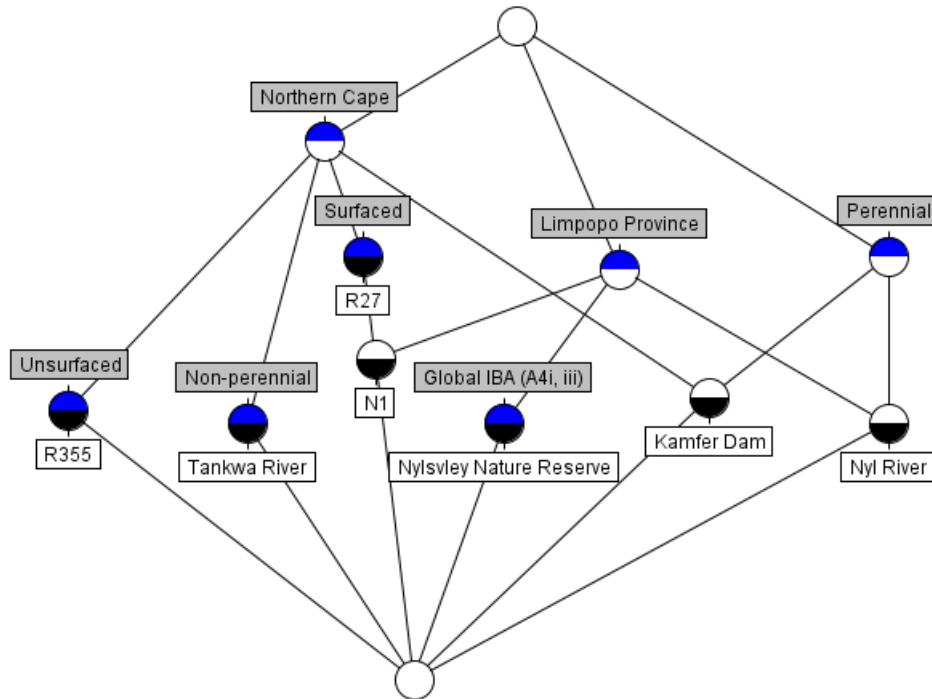


Figure 9.8: Feature instances in FCA

*geographical information* in particular. The UGC taxonomies of Wunsch-Vincent & Vickery [2007]; Gervais [2009]; Coleman *et al* [2009]; Budhathoki *et al* [2010]; Castelein *et al* [2010] and the typology of citizen science from Wiggins & Crowston [2011] have been used to build formal concept lattices (see Section 8.4 for details of the taxonomies). The classes in each of these taxonomies are the attributes in the concept lattice. For example, Gervais [2009] provides four classes for copyright issues, extending on the taxonomy in Wunsch-Vincent & Vickery [2007], namely *User-authored content*, *User-derived content*, *User-copied content* and *Peer-to-peer as UGC* (see Figure 9.10).

As discussed in Section 9.2.1, a formal context can be *one-valued* or *many-valued*. The taxonomies assessed here are generally *many-valued contexts*, but often without the domains of valid values being specified. It can be confusing and meaningless to use such attributes without their values in FCA. While an attribute such as *user-authored content* could be interpreted by the reader as meaning that an object (UGC repository) having this attribute in its extent was authored by the user (contributor), for example, an attribute such as *distribution platform* cannot be so interpreted. Further, having the attribute *quality* does not say if the quality is good or bad. Hence, it has been necessary to add attribute values in several instances for the analysis in Section 9.6. Hopefully, these additions are appropriate.

## 9. Using formal concept analysis to assess taxonomies

For the analysis reported on here, the objects are repositories of user-generated content (not necessarily ones available on the Internet). Coleman *et al* [2009] provide examples of such repositories (eg: *in-car navigation*, or an *open repository*), which I have supplemented, based on my experience, with two types: a *traditional SDI with strict control over its data sources* and *revision requests or notices submitted to an SDI*, see Figure 9.11. However, for the analysis in Section 9.6, the focus is primarily on the ten repositories identified in Section 8.3.2. The assignment of attributes to objects is based on the discussions of their taxonomies by Wunsch-Vincent & Vickery [2007]; Gervais [2009]; Coleman *et al* [2009]; Budhathoki *et al* [2010]; Castelein *et al* [2010], and my judgement see Sections 8.4 to 8.6.

Previously, Kokla & Kavouras [2001] used FCA to analyse the commonalities and differences between three “ontologies” (really, taxonomies) in geographical information, specifically, the way they classified the feature type *stream-watercourse*. Essentially (though they do not state it as such), Kokla & Kavouras [2001] used FCA to link up the sub-categories<sup>5</sup> from the three ontologies to produce a many-valued context, which they converted to a one-valued context to provide new, combined categories.

### 9.4.1 Discrimination adequacy

As mentioned above in Sections 9.2.2, 9.2.6 and 9.2.7, through attribute analysis, FCA can assist in identifying “missing” and “redundant” attributes and objects. This is then used here to determine the adequacy of taxonomies for discriminating between repositories containing UGC in general, or VGI in particular. They are illustrated using the theoretical example provided in Figure 9.9, to which the following subsections refer. Figure 9.9 shows absent and redundant attributes and objects. If a concept has own objects, they are shown below the node and if a concept has own attributes, they are shown above the node. Specifics of the *discrimination adequacy* of each of the selected taxonomies is shown in Figures 9.10 to 9.20.

For this analysis, I would suggest that a formal concept with low stability is generally more interesting. While this implies noise in the typical applications of FCA, in assessing the discrimination adequacy of a taxonomy, the low stability is due a higher prevalence of unique objects and/or attributes — it is more appropriate to consider these to be extreme values, rather than noise. The implications of high stability, and of missing and redundant attributes and objects, are discussed below in terms of the discrimination adequacy of pre-existing taxonomies of UGC when used to classify a target set of repositories of VGI.

- Are there formal concepts with few or no formal attributes in their intent? This would indicate repositories that are not classified by the taxonomy, and hence classes (formal attributes) that are missing.
- Are there formal concepts with few or no formal objects in their extent? This would indicate classes in the taxonomy for which there are few or no repositories, and hence repositories (formal objects) that are missing.

<sup>5</sup>Which they termed *semantic factors* as they were created by analysing and decomposing the categories semantically.

## 9. Using formal concept analysis to assess taxonomies

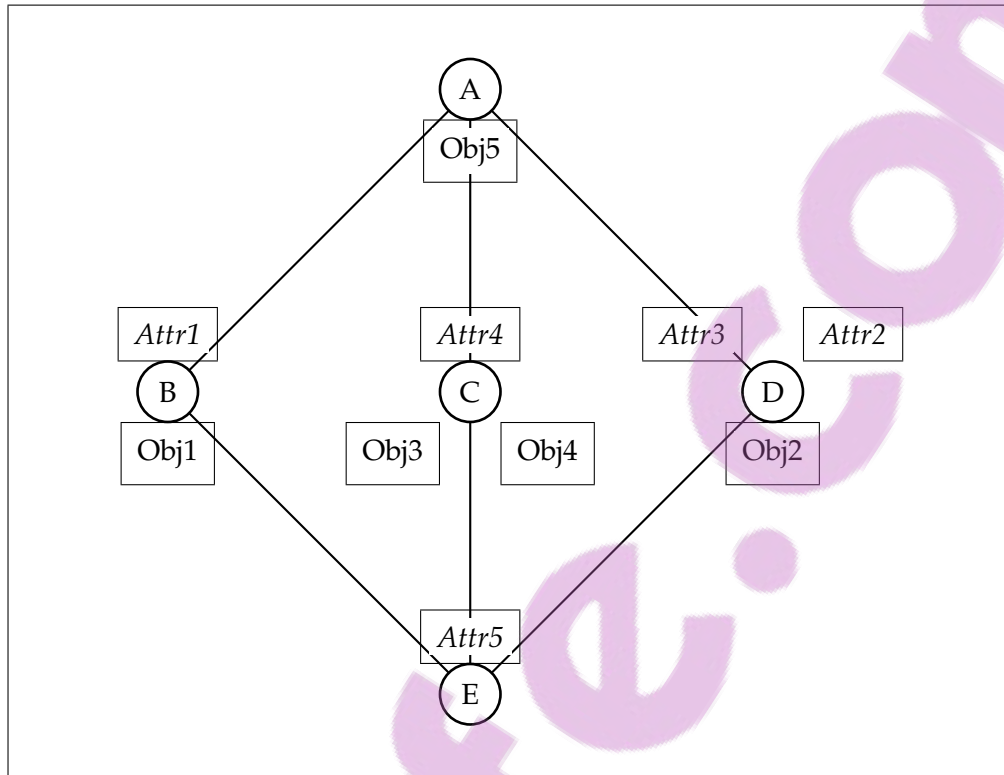


Figure 9.9: Absent and redundant attributes and objects.

- Are there attributes that are **redundant**? This would indicate taxonomy classes that are not differentiated from one another, which could be because they are redundant or poorly defined.

In my analysis, I expect to get **redundant** objects, because in the real world, there are likely to be several **such** repositories with identical classifications. The implications of these are discussed below in Section 9.6.

### 9.4.2 High **intensional** stability

Concept  $A = (\{Obj1, Obj2, Obj3, Obj4, Obj5\}, \emptyset)$  has high intensional stability, with its  $\sigma_i(A) = 0.84$ . Many (27) of the 32 subsets of  $A$ 's extent yield a concept whose intent is also  $\emptyset$ . The object  $Obj5$  is not described or differentiated by any attributes. For our analysis, this could be addressed by adding classes to a taxonomy, so that it can differentiate better between the repositories. The taxonomy in Gervais [2009] extends the taxonomy in Wunsch-Vincent & Vickery [2007] (namely, *Distribution platform* and *Type*), to cater for copyright issues. Without the classes of Gervais [2009], the taxonomy of Wunsch-Vincent & Vickery [2007] does not really differentiate repositories from one another.

Concept  $C = (\{Obj3, Obj4\}, \{Attr4\})$  also has relatively high intensional stability, with  $\sigma_i(C) = 0.75$ . The objects  $Obj3$  and  $Obj4$  are not differentiated from one another by the

## 9. Using formal concept analysis to assess taxonomies

attributes. Effectively,  $Obj3 = Obj4$  and one of them is redundant. In a comprehensive analysis of UGC repositories, one would expect this, namely repositories that are equivalent and hence direct competitors of one another. For example, referring to Figure 9.12 (which illustrates the taxonomy of Coleman *et al* [2009]), the objects shown are generic and there could be many repositories that are specific instances of each. Adding these repositories as objects would create redundancies in the taxonomy.

### 9.4.3 High extensional stability

Concept  $E = (\emptyset, \{Attr1, Attr2, Attr3, Attr4, Attr5\})$  has high extensional stability, with  $\sigma_i(E) = 0.84$ . Again, 27 of the 32 subsets of  $E$ 's intent yield a concept whose extent is also  $\emptyset$ . The attribute  $Attr5$  does not describe or differentiate any objects. This could be a weakness in the analysis, with an important type of repository omitted, or it could indicate a type of repository that does not yet exist and hence a potential “gap” in the market. While experimenting with FCA and the taxonomy of Budhathoki *et al* [2010], I realised the value of instability in a lattice. It highlighted a potential “gap” in the market, namely repositories that do not cater adequately for privacy: a widespread problem on the Internet.

Concept  $D = (\{Obj2\}, \{Attr2, Attr3\})$  also has relatively high extensional stability, with  $\sigma_e(D) = 0.75$ . No objects are differentiated from one another by the attributes  $Attr2$  and  $Attr3$ . Effectively,  $Attr2 = Attr3$  and one of them is redundant. This could be coincidental, could reflect a set of objects that is too narrow (eg: other types of repositories should also have been included), or could indicate that some classes should be removed from the taxonomy because they add no value or even worse, could cause confusion as users try to differentiate between classes that are, in essence, equivalent. We illustrate this in Figure 9.12 with a subset of the VGI taxonomy from Coleman *et al* [2009], for assessing the nature and motivation of *producers* (users who are also producers). For the objects, we use the generic examples of VGI repositories given by Coleman *et al* [2009]. As can be seen, there are several redundancies in the attributes, because these classes are inadequately defined, or cannot be differentiated in practice, or other types of repositories should be included in this analysis.

### 9.4.4 Missing formal attribute

A *missing formal attribute* would occur when there are two or more objects in an extent that one would expect to be differentiated from one another by their attributes, but they are not. The *extensional stability* would be *low* and FCA would highlight where this attribute is missing. The problem could be addressed by defining one or more suitable attributes for the intent of these objects, to separate them. In the context of the analysis presented here, this would involve adding one or more classes to a taxonomy, so that the taxonomy would differentiate better between the repositories. In other words, the taxonomy would be improved by adding the class.

The taxonomy in Gervais [2009] is an extension of the taxonomy in Wunsch-Vincent &

## 9. Using formal concept analysis to assess taxonomies

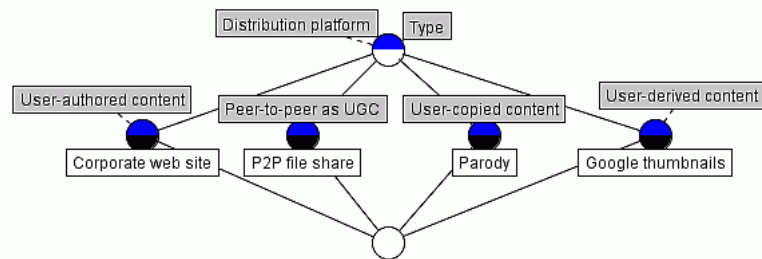


Figure 9.10: The taxonomy of Gervais [2009] for copyright issues for UGC.

Vickery [2007], to enable the latter to cater for copyright issues. Figure 9.10 shows the four classes for copyright issues added to the two main axes of the taxonomy in Wunsch-Vincent & Vickery [2007], which were *Distribution platform* and *Type*. As can be seen, removing the four classes for copyright issues would collapse the lattice into a very stable form, such as is shown in Figure 9.3.

### 9.4.5 Redundant formal attribute

Where *redundant formal attributes* occur, this could be coincidental, could reflect a set of objects that is too narrow (eg: there are other types of VGI repositories that should have been included in the analysis), or could indicate that some classes should be removed from the taxonomy because they add no value or even worse, could cause confusion as users try to differentiate between taxonomy classes that are, in essence, equivalent. From the FCA perspective, the *extensional stability* is too *high*.

We illustrate this with a subset of the taxonomy developed by Coleman *et al* [2009], for assessing the nature and motivation of *producers* (that is, users who are also producers), namely their five categories of expertise and four contexts for contributing VGI. Coleman *et al* [2009] do acknowledge that their categories overlap. For the objects, we use the generic examples given by Coleman *et al* [2009] in their Table 1, namely in-car navigation (eg: Tom Tom, Tele Atlas or HERE), open repository (eg: OpenStreetMap), public participatory geographical information system (PPGIS) and disaster reporting (eg: during the recent Haiti earthquake). As this set is a bit limited, we add to this two more types of repositories: a traditional SDI with strict control over its data sources, and revision requests or notices submitted to an SDI (eg: as described by Gu  lat [2009] for *swisstopo*, the national mapping agency in Switzerland).

Figure 9.11 shows the line diagram of this concept lattice. As can be seen, in this set of objects (spatial data repositories) and attributes (taxonomy classes), there is redundancy in the attributes *Interested Amateur* and *Expert Amateur*. This means that these two classes are inadequately defined, or cannot be differentiated in practice, or other types of repositories should be included in this analysis. In this case, the problem appears to be with differentiating between the classes in practice, because both interested and expert amateurs are likely to make the same types of contributions of UGC to the same types of

## 9. Using formal concept analysis to assess taxonomies

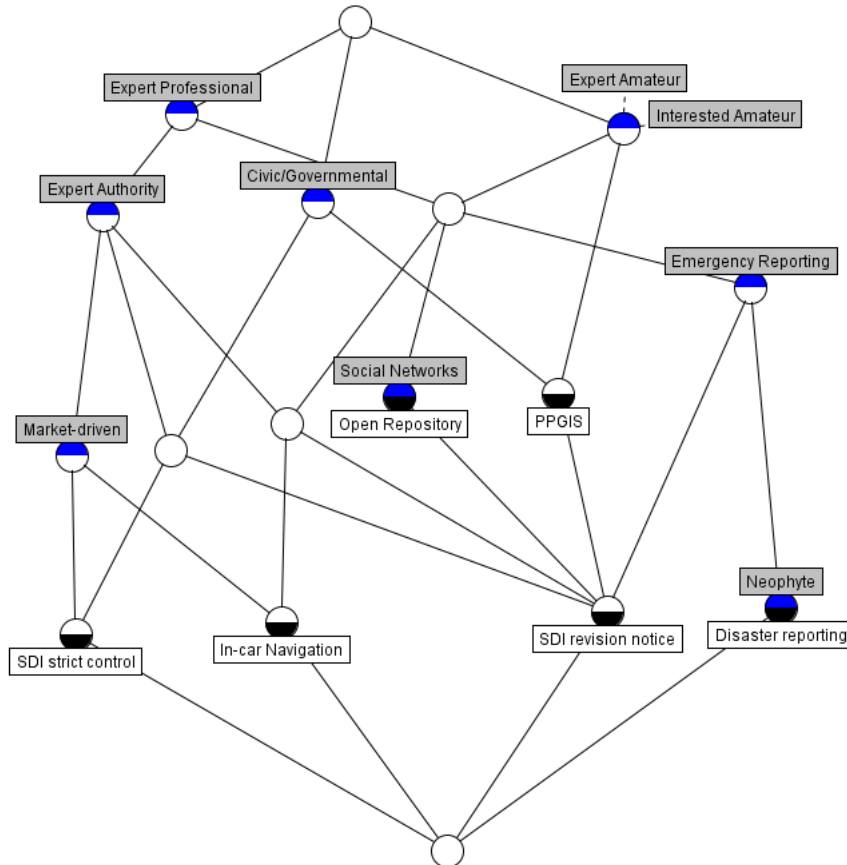


Figure 9.11: A subset of the taxonomy of Coleman *et al* [2009] for assessing the nature and motivation of *producers*.

repositories. Further, Figure 9.12 shows the lattice in Figure 9.11, but with just the repositories given in Table 1 of Coleman *et al* [2009]. One can see clearly the increase in the redundancy of the attributes.

### 9.4.6 Missing formal object

A missing object would be a type of VGI repository that has not been included in the analysis. This could be a weakness in the analysis, in that an important type of VGI repository had been omitted. Alternatively, it could indicate a type of VGI repository that does not yet exist and hence a potential “gap” in the market — revealed because of the *low intensional stability*. In a comprehensive analysis of all VGI repositories one would expect to find redundant objects, that is, VGI repositories that are fundamentally equivalent and hence direct competitors of one another, though possibly targeting different domains

## 9. Using formal concept analysis to assess taxonomies

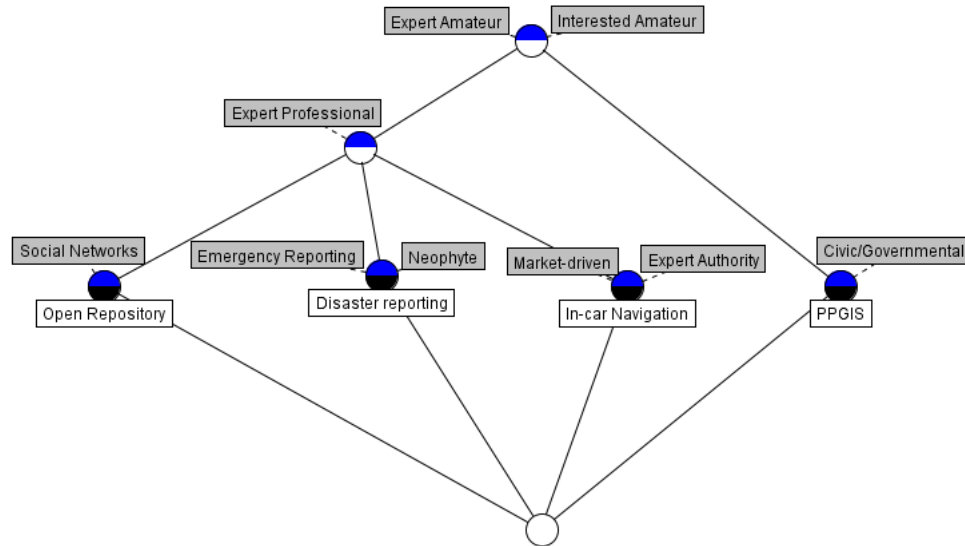


Figure 9.12: Figure 9.11 with only the repositories in Coleman *et al* [2009].

(assuming that the taxonomies do not differentiate on the domains). As I have used only a representative selection of VGI repositories, I do not expect many redundant objects.

It was while experimenting with FCA and the taxonomy of Budhathoki *et al* [2010], that I first discovered the value of instability in a lattice. This is to some extent an artificial example as it was not meant to be a definitive analysis of the repositories. However, as shown in Figure 9.6, it does illustrate a potential “gap” in the market — in this case, it would appear that repositories do not cater adequately for privacy, a widespread problem on the Internet<sup>6</sup>.

### 9.4.7 Redundant formal object

In a comprehensive analysis of repositories of UGC, one would expect to find redundant objects, that is, repositories that are fundamentally equivalent and hence direct competitors of one another, though possibly targeting different domains (assuming that the taxonomies do not differentiate on the domains). For example, referring to Figure 9.11, the objects are generic or abstract, and there could be many repositories that are specific instances of each. Adding these repositories as objects would create redundancies. From the FCA perspective, this would be high extensional stability.

<sup>6</sup>For example, Balkin [2014] and Zittrain [2014a] propose that online service providers should be *information fiduciaries* (as professionals such as doctors are expected to be with the information about their patients), because their users trust them with sensitive personal information which these companies can abuse.

## 9.5 Stability exploration

### 9.5.1 Background

While high intensional stability can reveal absent attributes or redundant objects, it cannot differentiate between absent attributes on the one hand, and redundant objects on the other. This is because qualitative analysis (ie: the insight of an expert) is required to differentiate between them. For example, referring to Figure 9.9 and Section 9.4.2, concept *C* has high intensional stability, but should *Obj3* or *Obj4* be removed, or some attribute added? This can only be determined by a domain expert.

Similarly, while a high extensional stability can reveal absent objects or redundant attributes, qualitative analysis is required to differentiate between absent objects and redundant attributes. Referring to Figure 9.9 and Section 9.4.3, concept *D* has high extensional stability, but should *Attr2* or *Attr3* be removed, or some object added? Again, this needs an expert to decide.

I would suggest that when assessing the discrimination adequacy of a taxonomy, the absence or redundancy of objects or attributes is undesirable. Such a taxonomy could be improved by reducing the intensional and/or extensional stability. It appears that this could be done in a manner similar to attribute exploration, as is done in ConExp [Yevtushenko *et al* 2003], starting with the concept with the “highest” stability (however that might be determined, as explained below) and then moving on to the next highest until the stability has been reduced appropriately. This process could be termed *stability exploration* [Cooper *et al* 2010b].

### 9.5.2 The rationale for stability exploration

In attribute exploration, adding a new object to satisfy a dependency that does not hold (i.e. through providing a counterexample) actually adds a new concept  $(A, B)$ , because it also introduces a new intent, as there was no other collection of objects before that had such an intent. The intensional stability of the new concept is therefore  $\frac{2^{|\zeta_i(A)|}}{2^{|A|}} = \frac{1}{2^{(|A|-1)}}$ , that is, the minimum possible value, when  $|\zeta_i(A)| = 1$ , as shown by Lemma 4 above, in Section 9.2.5. Similarly, in object exploration (ie: attribute exploration on a transposed context), adding a new attribute will add a new concept  $(A, B)$  with an extensional stability of  $\frac{2^{|\zeta_e(B)|}}{2^{|B|}} = \frac{1}{2^{(|B|-1)}}$ .

From here on, I will discuss only intensional stability to reduce confusion, but because of the dual relationship, as shown by Lemma 1, this all applies to extensional stability as well. Stability exploration based on minimizing the intensional stability alone can *remove* objects until there is only one left whose intent retains the intent of an initial lattice concept. In this sense, attribute/object exploration and stability exploration approach things from opposite sides.

However, stability exploration is not necessarily only for reducing stability. Hence, for some forms of stability exploration, the size of the stability *value* might be more important

## 9. Using formal concept analysis to assess taxonomies

---

than the normalized stability index for identifying the candidate objects or attributes to be removed or added, depending on the purpose of the exploration. For example, if two concepts each have the same stability index of  $\alpha$ , but one represents 192 different subsets out of 256 while the other represents 3 out of 4, one would probably rather trim down on the 192 subsets than on the 3.

In terms of the example in Figure 9.4, both  $\{Alice\}$  and  $\{Amy\}$  are  $\{Tall, Tanned, Young, Lovely\}$ . If it is important for us to have in the lattice at least one representative of those who are  $\{Tall, Tanned, Young, Lovely\}$ <sup>7</sup>, then we can remove at most one of the ladies to reach minimum stability. If it was important for us to have a better representation of  $\{Tall, Tanned, Young, Lovely\}$  people, even if they were to be  $\{Fat\}$  (a combination as yet unrepresented in the sample), then stability exploration would seek to increase the intentional stability of the relevant concept by asking the user if there were other examples of such people. In such a case, we would be indifferent as to which additional attributes the person had, whether  $\{Fat\}$  and/or  $\{Big\}$  and/or  $\{Bright\}$ , or indeed none of these. Attribute exploration, by contrast, would enquire whether a specific combination of existing attributes were to be found in a person, eg:  $\{Tall, Tanned, Young, Lovely \text{ AND } Fat\}$ .

Algorithmically, it hence seems that one should recompute the stability each time an action is taken, not only for the concept under scrutiny, but for all the other affected concepts as well. Further, it would be useful to specify automatic termination conditions for the stability exploration loop. For example, if one was aiming at stability minimization, such a condition could occur when a concept is at its lowest possible stability bound.

As stability exploration can result in the removal of objects or attributes (as well as adding them), there is a risk that for a concept selected for potential stability improvement, removing an object in its extent could effectively remove an attribute that is not in the concept's intent (in practice, the attribute would move to the intent of the zero). This is because the derivative of a object can be larger than its intent. Stability has already been used to prune a lattice automatically (eg: Roth *et al* [2006]), but the selection of objects or attributes to prune is specific to the application. With *stability exploration*, our intention is to develop a general algorithm to prune a lattice in an interactive process, as is done in attribute exploration. Both the special case of Roth *et al* [2006] and the general case of stability exploration require a user with expertise in the application domain.

### 9.5.3 A proposed methodology for stability exploration

I propose here a methodology for implementing stability exploration, to exploit the intensional and extensional stability values and indexes of a lattice to identify for which concepts the indexes are high and/or low, and hence where attributes and/or objects could be absent and/or redundant. As with attribute exploration, stability exploration would be an interactive process.

1. Sort concepts based on their stability value or index. The sort could be ascending or descending and it could be on the intensional or the extensional stability value

---

<sup>7</sup>A reasonable requirement!

## 9. Using formal concept analysis to assess taxonomies

or index, the sum of them, the product of them, the difference between them, etc. Hence, it is appropriate to give this sort index a specific name, which I shall term the *stability exploration index*.

2. In turn, present the first concept in the list to the user as a candidate for exhibiting redundancy or absence. The user can then identify redundant or absent attributes or objects and then act appropriately to improve the stability (either increase or decrease), as explained below, or do nothing.
  - (a) If the user identifies redundant attributes, remove attributes as the user considers appropriate, and repeat. Alternatively, the user could correct errors in the attributes, such as in the definitions.
  - (b) If the user identifies redundant objects, remove objects as the user considers appropriate, and repeat. Alternatively, the user could correct errors in the objects, such as in the definitions.
  - (c) If the user identifies absent attributes, add attributes as the user considers appropriate, and repeat. Alternatively, the user could correct errors in the attributes, such as in the definitions.
  - (d) If the user identifies absent objects, add objects as the user considers appropriate, and repeat. Alternatively, the user could correct errors in the objects, such as in the definitions.
3. Recompute the intensional and extensional stability values and re-sort the concepts.
4. Go to the next concept and repeat the process, or exit if enough has been done. Please note that after the stability values have been recomputed, the user could be presented again with the concept they have just worked on, so it might be better to present the next most suitable concept for editing. Further, the user could be presented with a list of all the concepts and their relevant stability values, so that the user can use their judgement to select the next concept for editing. Please note that the stability exploration could be stopped automatically when each concept has at most one own attribute and one own object. An interesting research question is whether or not the stability exploration could be stopped automatically at an earlier stage, such as when the change in the stability value is below some threshold?

For example, referring to Figure 9.4, a user could sort the concepts by their extensional stability indexes and present them from highest to lowest. The first concept presented will then be  $(\{Alice, Bob, Amy\}, \{Tall, Young\})$ , as it is the only concept with two own attributes ( $\xi_e = 2$ ). This is either because one of the attributes is redundant (ie:  $\{Tall\}$  and  $\{Young\}$  are synonyms), or because there is an absent object that is either  $\{Tall\}$  or  $\{Young\}$ , but not both. If the former is the problem, the user could remove either  $\{Tall\}$  or  $\{Young\}$ , but if the latter is the problem, the user could add another object, say  $\{Carol\}$ , with a derivative of either  $\{Tall\}$  or  $\{Young\}$ , but not both.

## 9. Using formal concept analysis to assess taxonomies

---

### 9.5.4 Possible applications of stability exploration

The following are examples of how stability exploration could be applied in practice. Please note that stability exploration is done using the *stability exploration index*.

- Stability exploration can be used as a *decision support tool*, such as for multi-criteria decision analysis (MCDA). Both FCA in general and stability exploration in particular are suited to the qualitative nature of MCDA. The objects in the lattice would be the alternatives and the attributes would be either the criteria for ranking or selecting the alternatives, or the characteristics of the alternatives.
- As proposed in Cooper *et al* [2010b], stability exploration can be used to assess the *discrimination adequacy* of taxonomies in other domains, such as bloodstain pattern analysis [Cooper 2003] (see Section 2.4.5.2 for some discussion on BPA taxonomies).
- Given a set of things to classify, stability exploration can be used to *build a taxonomy* for them.
- As a refinement of the previous bullet, given a population of things to be classified, stability exploration can be used to *assemble an hierarchical taxonomy* from a *data dictionary*, that is, from a flat taxonomy<sup>8</sup>.

## 9.6 Assessing the discrimination adequacy of existing taxonomies of UGC

A qualitative assessment of the five UGC taxonomies of Wunsch-Vincent & Vickery [2007]; Gervais [2009]; Coleman *et al* [2009]; Budhathoki *et al* [2010]; Castelein *et al* [2010] and the citizen science taxonomy of Wiggins & Crowston [2011] was done above in Sections 8.4 to 8.6, using the ten taxonomies described in Section 8.3.2. This assessment is taken further in this chapter, using formal concept analysis. Some of the analysis here matches that done in Chapter 8, while some is complementary. For example, Tables 8.3 to 8.8 give more details of how the selected repositories are classified by the six taxonomies, but Figure 9.6 and Figures 9.10 to 9.20 show how the taxonomies differentiate between the repositories.

A preliminary version of the analysis presented here and in Sections 8.4 to 8.6 was published in Cooper *et al* [2012b] and has been expanded and enhanced here, such as by the addition of the analysis of the typology of citizen science of Wiggins & Crowston [2011]. Some of the FCA analysis of these taxonomies is illustrated above in Figures 9.6, 9.10, 9.11 and 9.12, and is discussed in Sections 9.2.7 and 9.4.

## 9. Using formal concept analysis to assess taxonomies

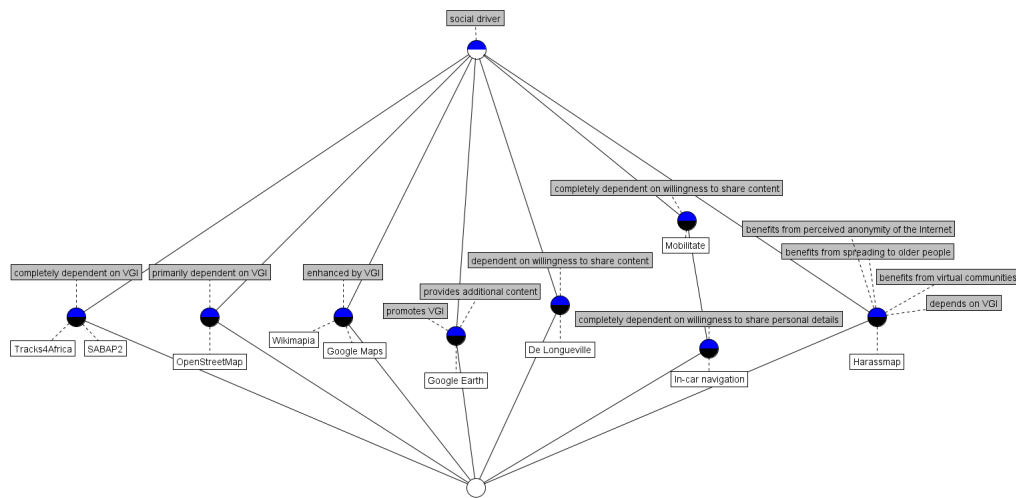


Figure 9.13: OECD's social drivers of UGC [Wunsch-Vincent & Vickery 2007].

### 9.6.1 OECD Working Party on the Information Economy

Table 8.3 above shows that the OECD's four drivers of user-created content [Wunsch-Vincent & Vickery 2007] are many-valued formal attributes, from an FCA perspective. For example, using conceptual scaling, the *social drivers* as identified in Table 8.3 could be transformed into the twelve one-valued formal attributes: *completely dependent on VGI*, *primarily dependent on VGI*, *depends on VGI*, *enhanced by VGI*, *provides additional content*, *promotes VGI*, *dependent on willingness to share content*, *completely dependent on willingness to share content*, *completely dependent on willingness to share personal details*, *benefits from virtual communities*, *benefits from perceived anonymity of the Internet* and *benefits from spreading to older people*. These attributes of the social drivers are shown in Figure 9.13, with the ten repositories.

On the other hand, the examples provided by Wunsch-Vincent & Vickery [2007] for the *social drivers* are also many-valued and could be transformed into the ten one-valued formal attributes: *shift to younger age groups (digital natives)*, *ICT skills*, *willingness to share content*, *willingness to share personal details*, *desire to create and express oneself*, *need for interactivity*, *development of virtual communities*, *development of collaborative projects*, *spread of these social drivers throughout older age groups* and *ability to fulfil societal functions (eg: social engagement)*. Obviously, there are similarities between these two sets of one-valued contexts, because I was applying the taxonomy of Wunsch-Vincent & Vickery [2007] in Table 8.3. I could also have used only the examples provided by Wunsch-Vincent & Vickery [2007]. However, these attributes are too vague to be used effectively in FCA to assess the *discrimination adequacy* of this taxonomy. While not perfect, the more refined attributes used for Table 8.3 and Figure 9.13 are better able discriminate between the repositories,

<sup>8</sup>Please note that Kokla & Kavouras [2001] recognised that hierarchical relationships could be *detected* when using FCA to compare ontologies.

## 9. Using formal concept analysis to assess taxonomies

as is shown by *completely dependent on VGI* vs *primarily dependent on VGI* vs *depends on VGI* vs *enhanced by VGI*, for example.

Clearly, adding in all the attributes in Table 8.3 for the other drivers, *technological*, *economic* and *institutional/legal*, to Figure 9.13 would discriminate completely between all ten repositories (making each an *own object*), but would add redundant attributes and increase the *intensional stability*. It would also make Figure 9.13 complex and difficult to interpret<sup>9</sup>, so it has not been done here. The redundant attributes could be weeded out or could be disaggregated by adding repositories, and this could be done manually, or by using *attribute*, *object* and/or *stability exploration*.

As can be seen, it can be difficult to use a taxonomy in FCA, unless its classes are well defined, logical and different from one another, as discussed in Section 2.4.

### 9.6.2 Gervais' taxonomy for copyright issues

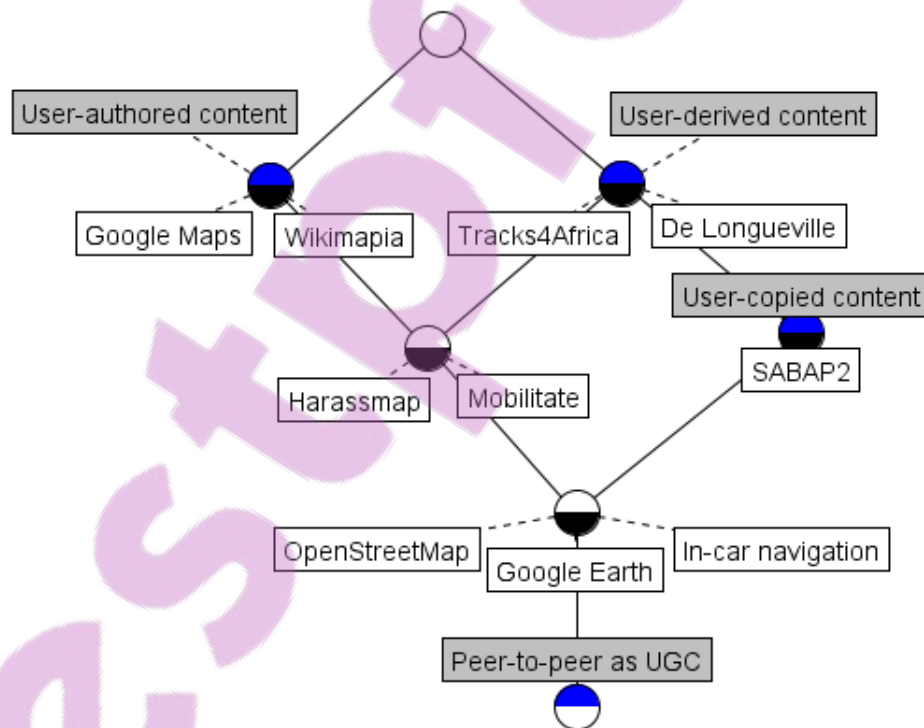


Figure 9.14: Copyright issues, from Gervais [2009].

<sup>9</sup>As it is, it is difficult enough to read!

## 9. Using formal concept analysis to assess taxonomies

Figures 9.10 and 9.14 show the taxonomy for copyright issues from Gervais [2009], which is an extension to the OECD taxonomy [Wunsch-Vincent & Vickery 2007]. The taxonomy is for user-generated content in general and while any repository could include content from all four of his classes, in practice it does show some discrimination between the ten selected repositories of VGI, as shown in Figure 9.14. This figure also shows that all the repositories have *user-authored* and/or *user-derived content*, and hence have VGI. None of the repositories have *peer-to-peer as UGC*, that is, unauthorized file sharing, which can also be seen in Table 8.4.

Figure 9.10 shows how the taxonomy discriminates between four generic types of UGC, that I selected based on the discussion in Gervais [2009].

### 9.6.3 Budhathoki, Nedovic-Budic and Bruce's framework for VGI

Figure 9.6 above shows the taxonomy of Budhathoki *et al* [2010] being used for *attribute exploration*. As this figure reveals, the attribute exploration showed *privacy* on the zero of the lattice, meaning that it was not an attribute of any of the repositories being assessed at that time. While it might be “obvious” that privacy is not dealt with well by many repositories on the Internet, this result of the attribute exploration revealed to me the value of using FCA for assessing taxonomies.

Subsequently, I added the repository *Harassmap* to my analysis and as Figure 9.15 shows, *privacy* is a distinguishing characteristic of *Harassmap*: if the women of Cairo are to report sexual harassment, they need to know that their identities will be protected, to prevent retaliation. In creating Figure 9.15, I used the functionality of ConExp [Yevtushenko *et al* 2003] to display a sub-context, by selecting and deselecting manually the attributes from Table 8.5 to use in the lattice, to try to discriminate between all the repositories with as few classes as possible. In this case, the five attributes *value*, *copyright*, *credibility*, *privacy* (from *issues*) and *structure* (from *contributing mechanisms*) are sufficient to discriminate completely between the ten repositories of VGI.

This ability to play with the attributes and objects in ConExp is another benefit of using FCA for assessing the discrimination adequacy of a taxonomy. This functionality in ConExp is for examining sub-contexts [Yevtushenko *et al* 2003] and in the ConExp documentation, it is not considered as a function for weeding out redundant attributes (those that are not that useful in the context). This process of selecting and deselecting attributes also differs from attribute exploration because the latter retains all the attributes and invites the user to add objects to create new implications, to remove dependencies between attributes. ConExp also uses all the attributes and objects of the context (whether or not a sub-context has been selected) to calculate the implications and associations and to do attribute exploration.

In comparison, my use of attribute exploration reduces the number of classes used from the taxonomy being assessed, to try to get a minimum set of attributes to discriminate between all the repositories completely. It gets to the core of the taxonomy, and also makes the line diagrams less cluttered. The difference is that conventional attribute exploration retains all the attributes, while my technique removes attributes.

## 9. Using formal concept analysis to assess taxonomies

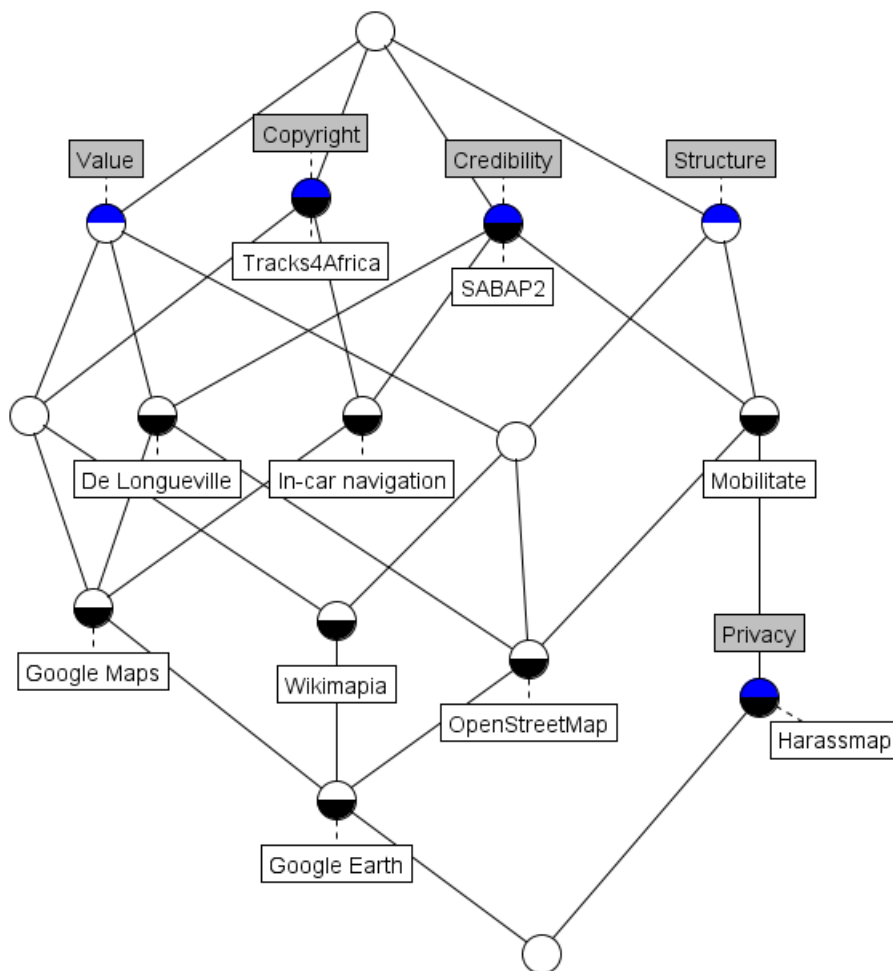


Figure 9.15: The ten taxonomies discriminated by only five attributes from Budhathoki *et al* [2010].

## 9. Using formal concept analysis to assess taxonomies

### 9.6.4 Coleman, Georgiadou and Labonte's nature and motivation of producers

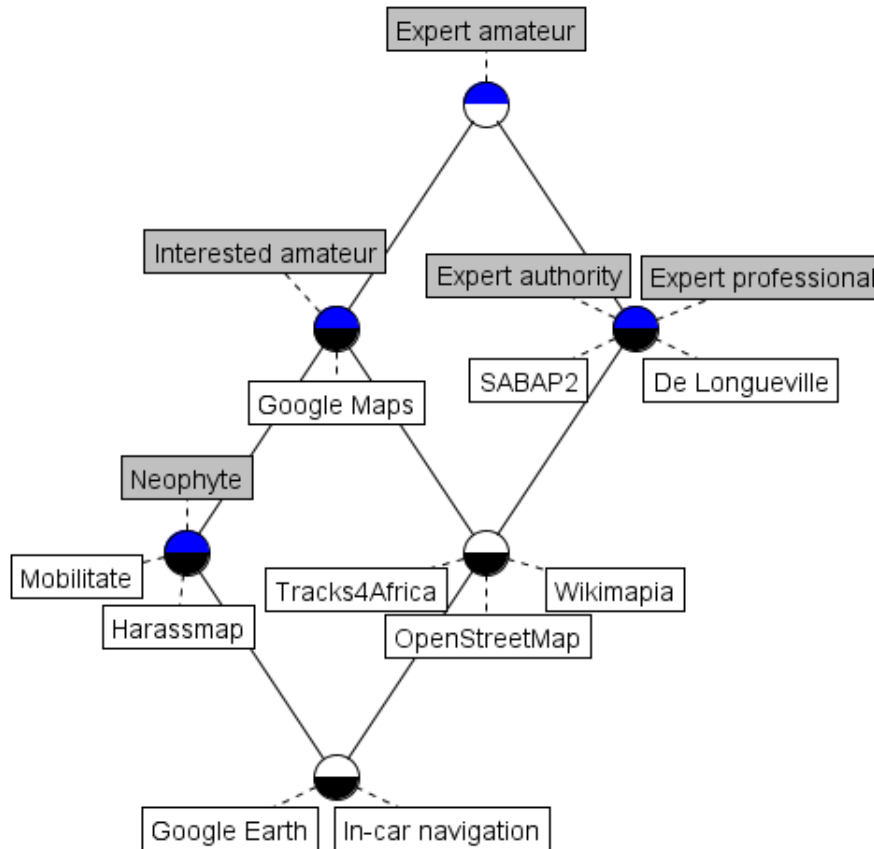


Figure 9.16: Expertise of producers [Coleman *et al* 2009].

The taxonomy of Coleman *et al* [2009] is illustrated above in Figures 9.11 and 9.12, which show their five categories of expertise and four contexts for contributing VGI as attributes, mapped against their four generic types of repositories, supplemented in Figure 9.11 with two generic types of repositories that I added. These figures were used to explain redundant formal attributes in FCA, see Section 9.4.5.

Figure 9.16 shows the same information as Table 8.6, namely the ten repositories classified by the five types of the expertise of producers, identified by Coleman *et al* [2009]. As can be seen, an *expert amateur* is an attribute of the unit of the lattice, as such a producer is likely to contribute to all of the different repositories. As discussed above in Section 9.4.5 and as is shown in Figure 9.16, an *interested amateur* and an *expert amateur* are likely to make the same types of contributions of VGI to the same types of repositories, so one of them is essentially redundant in this analysis. Nevertheless, it is probably very useful to be able to differentiate between the contributions of an *interested amateur* and an *expert*

## 9. Using formal concept analysis to assess taxonomies

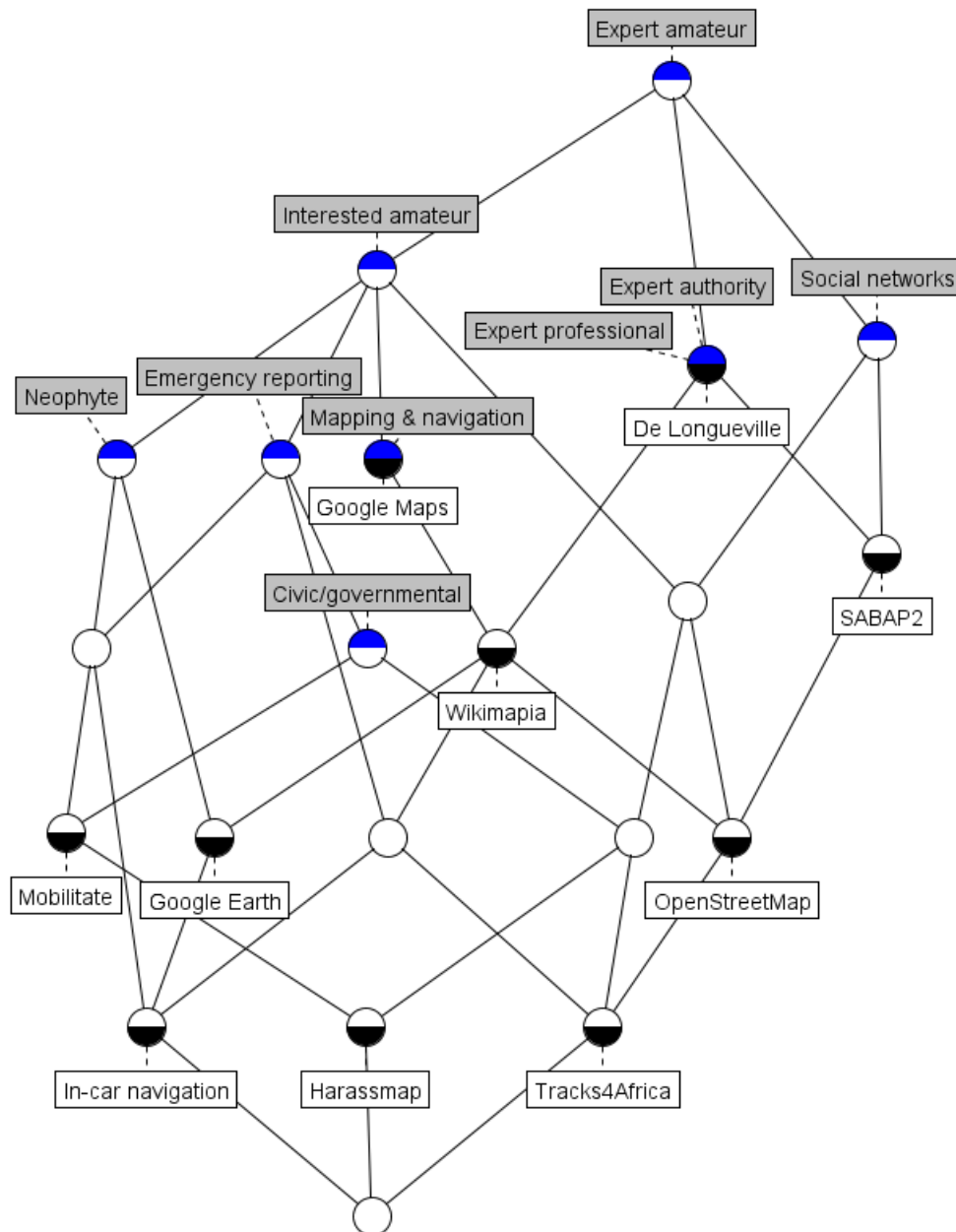


Figure 9.17: Expertise of producers and contexts for contributing [Coleman *et al* 2009].

## 9. Using formal concept analysis to assess taxonomies

*amateur*, so how could one define them separately and usefully?

This limited subset of all the characteristics identified by Coleman *et al* [2009] does not discriminate adequately between the ten repositories. However, Figure 9.17 shows the same data as Figure 9.16, but with the addition of the four generic types of repositories identified by Coleman *et al* [2009]. As can be seen, this subset of the characteristics identified by Coleman *et al* [2009] does discriminate completely between the ten repositories. Further, there is no overlap in this case between the four generic types of repositories of Coleman *et al* [2009]. However, these repository types still need the producer types to discriminate the repositories adequately: *neophyte* separates *Google Earth* from *Google Maps*, for example.

### 9.6.5 Castelein, Grus, Crompvoets and Bregt's characterization of repositories of VGI

Castelein *et al* [2010] identified thirteen characteristics to describe VGI in terms of the five SDI components in the conceptual model of Rajabifard *et al* [2002] (policy, access networks, technical standards, data and people), see Table 8.7. Unfortunately, as discussed in Section 8.4.6, four of these characteristics are numerical and really overlap, as they reflect the popularity of the SDI. Figure 9.18 shows this taxonomy without these numerical characteristics, with the five repositories that Castelein *et al* [2010] used as case studies classified according to Table 2 in Castelein *et al* [2010]. As can be seen, two of the classes are completely redundant for classifying these repositories: all the repositories allow *uploading* and all cater for *point data*. There are several other cases of redundancy, such as one of *polygons* and *feature types standard*.

Figure 9.19 shows the ten repositories discriminated completed by five of the characteristics of Castelein *et al* [2010]. Again, I used ConExp [Yevtushenko *et al* 2003] to select and deselect manually the characteristics of Castelein *et al* [2010] to use. Please note that SABAP2 is an own object of the zero of the lattice because all five attributes apply to it. For example, *specific thematic focus* separates SABAP2 from *Google Maps*, and *download* separates SABAP2 from *Mobilitate*.

### 9.6.6 Wiggins and Crowston's typology of citizen science

As opposed to the analysis of the five taxonomies of UGC/VGI given above, the typology of citizen science of Wiggins & Crowston [2011] was not included in Cooper *et al* [2012b].

Figure 9.20 shows the VGI repositories classified according to the typology of citizen science of Wiggins & Crowston [2011], modified by my addition of the typology *Subject*, as explained in Sections 4.4.2 and 8.5. As would be expected from Table 8.8, *Investigation* is an attribute of the unit of the context, and the repositories *Google Maps*, *Google Earth*, *Wikimapia* and *Tracks4Africa* are not differentiated from one another, as they are general repositories of VGI. The remaining six VGI repositories are identified uniquely by this modification of the typology of citizen science of Wiggins & Crowston [2011].

## 9. Using formal concept analysis to assess taxonomies

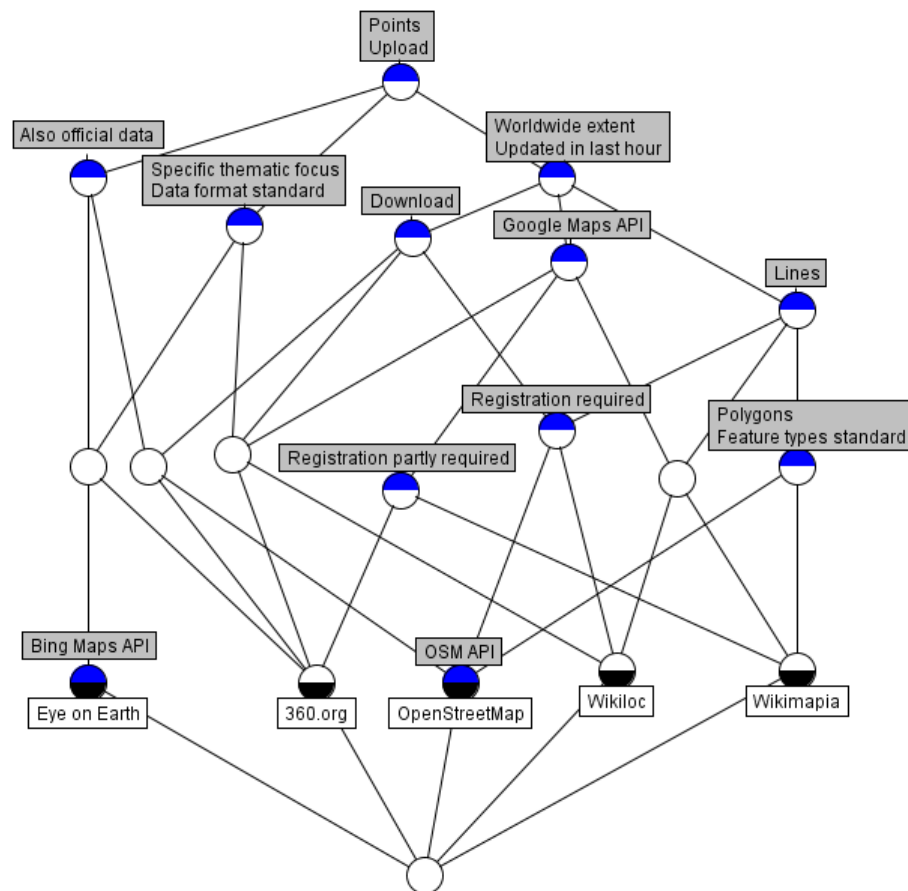


Figure 9.18: Characterization of VGI repositories [Castelein *et al* 2010]

## 9. Using formal concept analysis to assess taxonomies

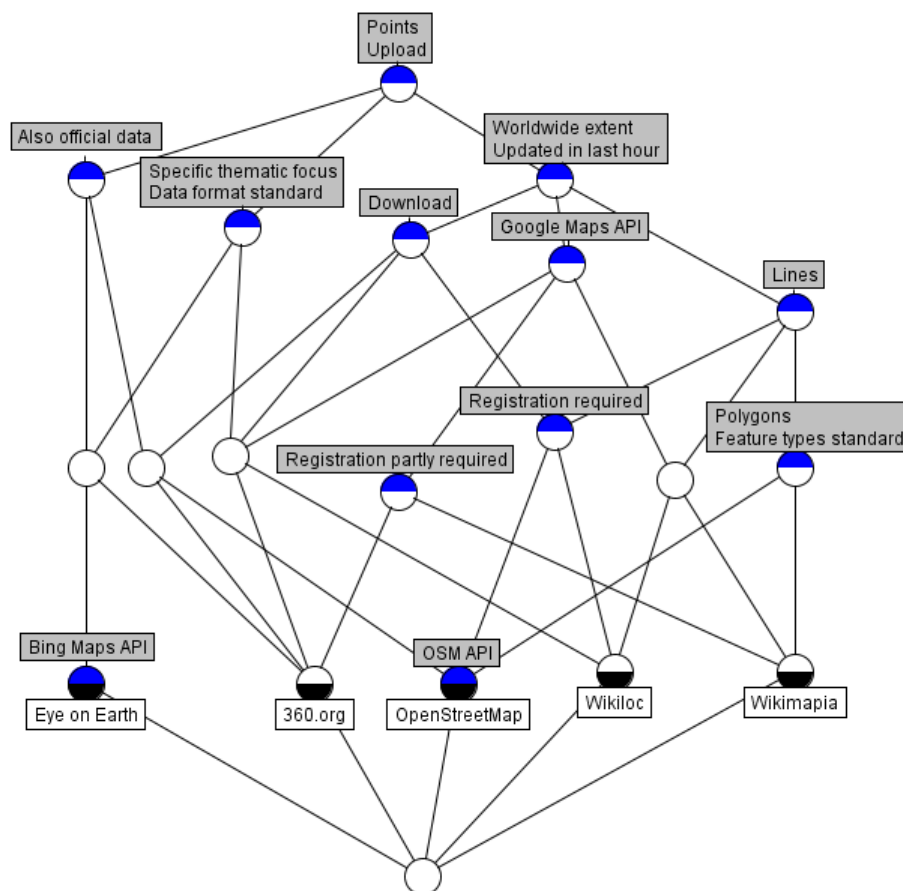


Figure 9.19: The ten taxonomies discriminated by only five attributes from Castelein *et al* [2010].

## 9. Using formal concept analysis to assess taxonomies

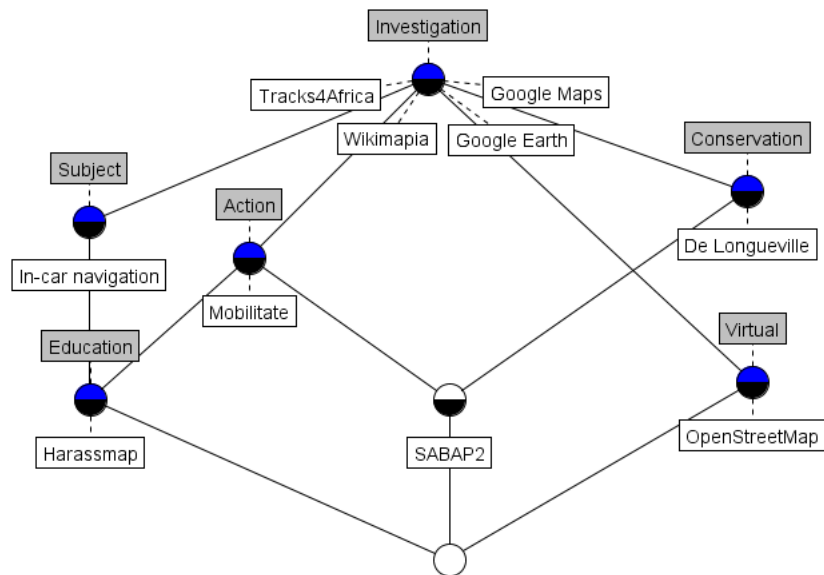


Figure 9.20: Wiggins and Crowston's typology of citizen science and VGI repositories

### 9.6.7 Commentary

User-generated content in general, and volunteered geographical information in particular, are becoming more important as data sources. Discovering and assessing the suitability of VGI for one's purposes is hence becoming more important, but can be difficult. One way of assessing VGI resources is by using a taxonomy [Cooper *et al* 2012b].

The five UGC/VGI taxonomies discussed above took different perspectives of UGC and VGI: Wunsch-Vincent & Vickery [2007] considered the drivers of UGC, Gervais [2009] the copyright issues, Budhathoki *et al* [2010] a framework for conceptualizing VGI, Coleman *et al* [2009] the nature and motivation of producers, and Castelein *et al* [2010] considered repositories of VGI from the perspective of the SDI components of Rajabifard *et al* [2002]. Wiggins & Crowston [2011] examined the nature of citizen science. These taxonomies considered the nature of UGC, the nature of the repositories of UGC, and the nature and motivations of the creators of UGC. Other factors that could be considered for such a taxonomy include the authority of the UGC, vested interests, intelligibility and quality of the UGC, ethics, liability and the availability of metadata [Cooper *et al* 2012b].

Sections 9.6.1 to 9.6.6 used formal concept analysis to illustrate these taxonomies and to assess their ability to discriminate between ten repositories of VGI, identified in Section 8.3.2. Sections 9.6.1 to 9.6.6 also compare the FCA analysis to the qualitative analysis of these six taxonomies, discussed above in Sections 8.4 and 8.6. They showed how the qualitative analysis and FCA analysis complement each other. While the tables in Section 8.4 might give more details of how the selected repositories are classified by the six

## 9. Using formal concept analysis to assess taxonomies

taxonomies, the figures in this chapter show how the taxonomies differentiate between the repositories, that is, the *discrimination adequacy* of these taxonomies. Further, using an FCA tool such as ConExp [Yevtushenko *et al* 2003] can highlight those attributes (ie: classes in a taxonomy) that are redundant or that are most effective, allowing one to trim down the taxonomy to its most useful classes.

Clearly, FCA can be used effectively to analyse a taxonomy.

### 9.6.8 Future research directions

The FCA analysis presented here could be expanded to assess each of these six taxonomies *in toto*, to investigate other repositories of VGI or UGC, or to assess a combination of the taxonomies. The following are other possibilities for future research on using FCA for assessing taxonomies.

- FCA could be used to assess taxonomies in **other domains**, such as to build on my previous work on bloodstain pattern analysis [Cooper 2003], see Section 2.4.5.2.
- Classifying and indexing geospatial resources, such as through the taxonomies discussed in this chapter and in Chapter 8, are needed for the discovery of geospatial resources, such as VGI. However, as discussed above (in particular, see Section 4.5), VGI is a new field of research and there are **differing interpretations of VGI** (eg: that of De Longueville *et al* [2010b], see Section 8.3.2.8), which makes such taxonomies even more essential [Cooper *et al* 2012b].
- Further research is needed into the nature of VGI before a comprehensive taxonomy of VGI can be produced. Any such taxonomy should probably be integrated with the **metadata** of the VGI data sets, preferably into a metadata standard such as ISO 19115 [2003], [Cooper *et al* 2012b].
- Finally, as alluded to in Section 9.2.7, the utility of **object exploration** could be studied.

## 9.7 Summary and looking ahead

This chapter has built on all the preceding chapters in this thesis to use formal concept analysis to present an assessment of various repositories of VGI and taxonomies of VGI. For this, the chapter presented an overview of FCA, introduced the concept of stability exploration within FCA, and presented some lemmas on stability in a lattice. In contrast to the usual applications of FCA, I have shown here that instability in a lattice can have value for analysis.

This chapter has also outlined the correspondence between the feature model and FCA; explored how FCA can be used to assess a taxonomy, covering discrimination adequacy, absent and redundant attributes and objects, and high intensional and extensional stability; and then used FCA to assess the discrimination adequacy of the taxonomies discussed in Chapter 8.

## *9. Using formal concept analysis to assess taxonomies*

---

The major original contributions that I have made that are presented in this chapter are:

- I determined how to use **formal concept analysis** for assessing existing taxonomies;
- In FCA, I determined that there can possibly be value in instability in a lattice when assessing a taxonomy;
- I contributed a few lemmas on stability in a lattice; and
- In FCA, I discovered stability exploration and developed a specification of it.

Further, the key contributions that I have made that are presented in this chapter are:

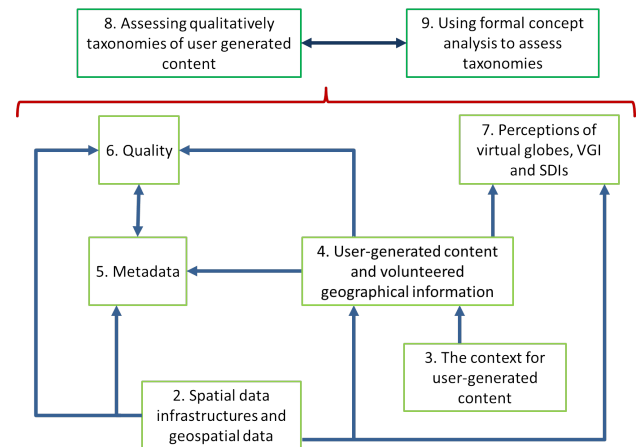
- The correlation between FCA and the feature model;
- An assessment of the discrimination adequacy of six taxonomies, using FCA; and
- Using an FCA tool to find the most effective combination of attributes.

Chapter 10 now provides the overall conclusions for my thesis, summarizing how this work facilitates integrating VGI into SDIs. It includes those questions for further research posed in various chapters. Chapter 10 is followed by the appendices form an integral part of this thesis. They have been placed at the end of this thesis for ease of reference.

\*\*\*\*

---

*9. Using formal concept analysis to assess taxonomies*



## Chapter 10

# Conclusions

### 10.1 Overview

This thesis and the research it covers have provided an exposition of the nature of volunteered geographical information (VGI) and its suitability for integration into spatial data infrastructures (SDIs). To do this, it analysed the nature of VGI and its applicability for use in an SDI to supplement official and commercial sources, particularly given the ease with which ordinary people can document their environment, experiences, perspectives and prejudices, share them widely and rapidly, and even query anyone else's content. For this research, taxonomies and repositories of such information were examined qualitatively and using formal concept analysis (FCA).

Further, this thesis attempts to reflect on the context for SDIs and VGI and the challenges and opportunities for both. For this, it first provided the context: terminology, geospatial data, inter networking, user-generated content (UGC), classification, folksonomies, citizen science, crowd sourcing, neogeography, metadata, quality, standards, the limitations of the Internet, and formal concept analysis.

VGI comes in many different forms of varying quality, and is interpreted in different

ways, as has been discussed herein. Hence, separating the good VGI from the bad is difficult. SDIs are an evolving concept, which means that the requirements for content suitable for an SDI are also changing. SDIs can make an important contribution to a country's development, such as through South Africa's National Development Plan. What is reported on here addresses only some of the problems of the SDI value chain, from content both professional and VGI, through to the SDI itself.

Nevertheless, this thesis has assessed perceptions of VGI, SDIs and virtual globes; it has conducted a qualitative assessment of the ability of various taxonomies of UGC and citizen science to discriminate adequately between repositories of VGI; and used FCA to conduct a more rigorous analysis of these taxonomies, covering discrimination adequacy, absent and redundant attributes and objects, and high intensional and extensional stability. In addition, this thesis includes lemmas on stability in a lattice (providing lower and upper bounds for intensional and extensional stability indices), original contributions on stability exploration and on the value of instability in a lattice.

The six taxonomies analysed took different perspectives of UGC and VGI: the drivers of UGC, the copyright issues, a framework for conceptualizing VGI, the nature and motivation of producers, repositories of VGI from the perspective of SDI components, and the nature of citizen science. These taxonomies considered the nature of UGC, the nature of the repositories of UGC, and the nature and motivations of the creators of UGC. The analysis using FCA showed how these taxonomies could discriminate between the repositories, and how FCA could be used to identify the classes within the taxonomies that are most effective. This research showed how the qualitative analysis and FCA analysis complement each other. While the qualitative tables might give more details of how the selected repositories are classified by the six taxonomies, the FCA figures show the discrimination adequacy of these taxonomies. The analysis also showed how these taxonomies could be improved.

Section 10.2 provides a review of this document and Section 10.3 lists some research questions raised herein.

## 10.2 Review

As an exposition of the nature of VGI and its suitability for integration into SDIs, this thesis provided the context, and then assessed taxonomies and repositories of VGI, both qualitatively and using FCA. The context is needed for understanding SDIs, VGI and how VGI can contribute to an SDI. The chapters providing the context also make important contributions as part of my research and this thesis.

Specifically, this thesis has discussed SDIs in South Africa and elsewhere; the terminology, types and complexities of geospatial data; classification, ontology and their encoding, including the curse of clever codes; models used in GISs; formal models of SDIs; data quality and metadata; incremental updating and versioning; cartography; and virtual globes and geobrowsers.

Then, this thesis provided details of the context that made the proliferation of UGC and

## 10. Conclusions

---

VGI possible, and the impact thereof: inter-networking, services, the Semantic Web, social mapping, (impossibility of) controlling the Internet, open archives and access, privacy, censorship, liability, patents, copyright, curation, the digital divide and standards.

As they can be confused with one another and to provide the context for VGI, this thesis provided details of UGC, citizen science, VGI, crowd sourcing and neogeography, together with the validity of using UGC in scholarly research, the quality of the traditional media, and citing UGC, data and data repositories.

Concerning metadata, this thesis has covered definitions; aspects, such as its relationship to quality; active and passive metadata; the benefits of encoding metadata; the relationship between a product specification and metadata; tools for capturing metadata; the different categories or types of metadata; principles for standards for metadata; selected standards; limitations of metadata; comparing searching to metadata; metadata and linked open data; and VGI and metadata.

Concerning quality, this thesis has provided details of the different aspects of the quality of resources, the four stages for recognising the quality of a resource, GNSS errors, the dimensions of quality, challenges for the quality of VGI, the quality of three VGI repositories, quality and classification, and standards for the quality of geospatial data.

This thesis has reported on the results of a survey conducted through a questionnaire, of geographical information professionals in Africa in general and in South Africa in particular, concerning their perceptions of virtual globes, VGI and SDIs.

Penultimately, this thesis has presented a qualitative assessment of various repositories of VGI and taxonomies of VGI, both separately and against one another. For this, it discussed the five taxonomies of UGC and VGI and one of citizen science used; the ten VGI repositories selected for assessment; the candidate repositories that were not selected; the qualitative assessment itself; and a preliminary taxonomy of UGC that I developed.

Finally, building on all the above, this thesis used FCA to present an assessment of various repositories of VGI and taxonomies of VGI. For this, it presented an overview of FCA, introduced the concept of stability exploration within FCA, and presented some lemmas on stability in a lattice. In contrast to the usual applications of FCA, I have shown here that instability in a lattice can have value for analysis. This thesis has also outlined the correspondence between the feature model and FCA; explored how FCA can be used to assess a taxonomy, covering discrimination adequacy, absent and redundant attributes and objects, and high intensional and extensional stability; and then used FCA to assess the discrimination adequacy of the taxonomies

This thesis has also presented several original research contributions, to information science, to geographical information science and to theoretical computer science:

- For FCA, it presents several lemmas on stability in a lattice (providing lower and upper bounds for intensional and extensional stability indices);
- It shows that there can be value in instability in a lattice when assessing a taxonomy (the instability represents extreme values rather than noise);
- It presents stability exploration, which could be used as a decision support tool;

- It describe the four stages for recognising the quality of a resource in general;
- It reports on a survey of geographical information professionals on VGI, SDIs and virtual globes;
- It clarifies the differences between UGC, VGI, citizen science, crowd sourcing and neogeography, which can be confused with one another; and
- This thesis explains why the Internet cannot be controlled.

### 10.3 Future research topics

Clearly, the qualitative analysis (see Chapter 8) and the FCA analysis (see Chapter 9) used in this thesis could be expanded to assess each of the taxonomies in greater detail, to investigate other repositories of VGI or UGC, to assess a combination of the taxonomies, and/or to assess taxonomies in other domains. Given the different interpretations of UGC, VGI, citizen science, crowd sourcing and neogeography, there needs to be more research on what they are and are not. Further, a number of research questions were raised in the preceding chapters, and they have been gathered here.

To what extent will citizens be prepared to adhere to policies and other standards, on which they invariably have had no input? How can citizens be included in the development of policies and standards, whether or not they are yet VGI contributors?

Can an expert hierarchical taxonomy provide more certainty than raw searching, or should one just dispense with taxonomies? What role does metadata play in actually enabling the linkages within linked data, and the linkages when integrating linked data? Can the concept of linked data be extended to metadata to create linked metadata, that is, linking items in metadata with one another?

Does the 1% rule make the Web more radical, because the 1% that are the creators of original content have very different perspectives from the 90% that are lurkers? Does the need to stand out from the silent majority and defend one's position encourage more extreme positions? Does this create and reinforce filter bubbles, as such unsparing attitudes discourage engagement and encourage users to seek out safe harbours where the opinions and declarations match their own perspectives?

How can SDIs, geovirtual environments and other repositories of geospatial data address information poverty and the digital divide? Do virtual globes and other repositories of VGI entrench or exacerbate the digital divide? How can geospatial services on mobile devices help users understand their spatial context and impact on others? Does too much bandwidth actually result in lower-quality VGI, effectively providing quantity rather than quality? How should a virtual globe decide how to prioritise the data that can be displayed?

How does one balance maintaining the integrity of VGI with making it easy for arbitrary producers of VGI to continue providing VGI, and to keep on improving quality? How can one ensure that the abstract model used for the geospatial data will enhance quality?

## 10. Conclusions

---

How does one improve the classification correctness, updating efficiency, completeness (particularly ensuring consistent coverage across the whole of the domain) and metadata for VGI?

Are virtual globes having any significant impact on official mapping now, for example, by encouraging mapping agencies to improve quality, availability and consumer orientation? Does the legislative and policy environment encourage the development of SDIs, and the development of, use of, and adherence to, standards, and encourage more than stifle innovation in the field of geographical information? Do the legislative and policy environments still deal poorly with issues such as virtual globes, VGI and open access to geographical information and require further research?

Referring to the levels of expertise of producers, is it possible to differentiate meaningfully between the contributions of an interested amateur and an expert amateur?

Could a taxonomy of VGI be integrated with the metadata of a VGI data set, perhaps using a metadata standard such as ISO 19115-1 [2014]?

Referring to the lemmas on stability in a lattice, is there a stability index value above which every concept is stable, and/or below which every concept is unstable, or does it depend on the application? Could stability exploration be stopped automatically before each concept has at most one own attribute and one own object, such as when the change in the stability value is below some threshold? Could a prototype stability exploration system be implemented that interfaces to an existing FCA tool, such as ConExp? What possible applications are there for object exploration?

\*\*\*\*

---

## *10. Conclusions*

## Appendix A

# Questionnaire on VGI

As discussed above in Chapter 7, the author designed a questionnaire to assess the understanding of virtual globes, geobrowsers and volunteered geographical information. The questions are included below, structured more or less as they were in the original questionnaire, together with their multiple choice answers or boxes for free-text answers, which ever might have been the case.

Table A.1:

### Questionnaire on virtual globes and geobrowsers

*This questionnaire has been compiled as a follow-up to a paper on geographical information perspectives on innovation and economic growth, to be presented at the first session of the Committee on Development Information, Science and Technology (CODIST) in Addis Ababa, Ethiopia, from 27 April to 1 May 2009.*

1. Your country of current residence						
2. In which sector of the economy are you employed?						
Government (national, provincial or local)	United Nations or related in- ternational agency	State-owned enterprise (eg: utility, science council)	Academia (including full-time students)	Private sector (including self- employed)	Non- government organisation (NGO)	Other (including retired, un- employed)

## A. Questionnaire on VGI

*A virtual globe provides masses of digital geographical information over the Internet, typically in the form of a globe.*

*A geobrowser is the interface to a virtual globe, typically allowing users to zoom into the data, switch data layers on and off, create three-dimensional views and add their own data (user generated content), such as geographical features (eg: roads and places of interest), tags (with text or links to web sites) and photographs.*

*Perhaps the best-known example of a virtual globe/geobrowser is Google Earth.*

3. What do you think is/are the main advantage(s) of virtual globes and geobrowsers?				
4. What do you think is/are the main disadvantage(s) of virtual globes and geobrowsers?				
5. What do you think is/are the main advantage(s) of user generated content in a virtual globe/geobrowser?				
6. What do you think is/are the main disadvantage(s) of user generated content in a virtual globe/geobrowser?				
7. What do you think of the documentation of the data (ie: the metadata) in virtual globes/geobrowsers?				
8. What do you think of the quality of the data in virtual globes/geobrowsers?				
9. What impacts are virtual globes/geobrowsers having now on the official mapping in your country?				
10. What impacts do you think virtual globes/geobrowsers will have on the official mapping in your country over the next five years (through to 31 December 2014)?				
11. Do you think that the legislative and policy environment in your country encourages or stifles innovation in the field of geographical information?				
Encourages innovation	Neither	Stifles innovation	Don't know	
12. Do you think that the legislative and policy environment in your country encourages or stifles the development of spatial data infrastructures (SDIs)?				
Encourages SDIs	Neither	Stifles SDIs	Don't know	
13. Do you think that the legislative and policy environment in your country encourages or stifles the development of, use of, and adherence to, standards?				
Encourages standards	Neither	Stifles standards	Don't know	
14. How well do you think the legislative and policy environment in your country deals with issues such as virtual globes, volunteered geographical information and open access to geographical information?				
Very well	Adequately	Poorly	Does not cater for them at all	Don't know



### A. Questionnaire on VGI

15. Do you have access to a virtual globe/geobrowser at home?	Yes	No
16. Do you have access to a virtual globe/geobrowser at work?	Yes	No
17. Do you use a virtual globe/geobrowser for personal use?	Yes	No
18. Do you use a virtual globe/geobrowser for work purposes?	Yes	No

*Other than the last question (concerning your contact details), the remaining questions are only relevant if you use a virtual globe/geobrowser.*

19. If you use a virtual globe/geobrowser, which ones do you use? You may select more than one.				
Google Earth	NASA World Wind	Open Street Map	Microsoft Virtual Earth	Yahoo! Maps
Other (please specify)				
20. What are the main reasons you use a virtual globe/geobrowser? You may select more than one.				
Travel planning (work or leisure)	Providing a geographical context to news items	Accessing data for work purposes	General curiosity	Publishing your data
Reconnaissance for work purposes	Providing a geographical context to correspondence from friends and family	Backdrop for other geographical data	Armchair travelling	Searching for data
Other (please specify)				

21. Do you use the user generated content (volunteered geographical information) in a virtual globe/geobrowser?	Yes	No
22. Do you use a markup language in a geobrowser, such as the Keyhole Markup Language (KML)?	Yes	No

*If you are interested in getting feedback on this survey or participating in follow-up surveys, please include your name and email address below (please write clearly!). If you would prefer your questionnaire response to remain anonymous, you can rather email your contact details to my address below.*

Family name	Given name or initials	Email address

**Thank you!**

*Please return to: Antony Cooper, Built Environment Unit, CSIR, PO Box 395, Pretoria, 0001, South Africa*

*Facsimile: +27 12 841 3037. Email: acooper@csir.co.za*

---

*A. Questionnaire on VGI*

---

## Appendix B

# Published taxonomies of user generated content

### B.1 Overview of the appendix

This appendix provides the details of the five taxonomies discussed in Chapter 8, for ease of reference. Wunsch-Vincent & Vickery [2007] is presented in Section B.2, Gervais [2009] in Section B.3, Budhathoki *et al* [2009] in Section B.4, Coleman *et al* [2009] in Section B.5 and Castelein *et al* [2010] in Section B.6.

### B.2 OECD Working Party on the Information Economy

Wunsch-Vincent & Vickery [2007] present the following four groups of drivers of user-created content, which have been taken verbatim from Box 1 in their paper, but are included here for ease of reference. They are discussed above in Section 8.4.2:

- *Technological Drivers*
  - Increased broadband availability
  - Increased hard drive capacity and processing speeds coupled with lower costs
  - Rise of technologies to create, distribute, and share content
  - Provision of simpler software tools for creating, editing, and remixing
  - Decrease in cost and increase in quality of consumer technology devices for audio, photo, and video
  - Rise of non-professional and professional UCC sites as outlets
- *Social Drivers*

## *B. Published taxonomies of user generated content*

---

- Shift to younger age groups (“digital natives”) with substantial ICT skills, willingness to engage online (ie: sharing content, recommending and rating content, etc.) and with less hesitation to reveal personal information online
- Desire to create and express oneself and need for more interactivity than on traditional media platforms such as TV
- Development of communities and collaborative projects
- Spread of these social drivers throughout older age groups and to fulfil certain societal functions (social engagement, politics and education)
- *Economic Drivers*
  - Lower costs and increased availability of tools for the creation of UCC (eg: for creating, editing, hosting content) and lower entry barriers
  - Increased possibilities to finance related ventures and UCC sites through venture capital and other investment possibilities
  - Lower cost of broadband Internet connections
  - Increased interest of commercial entities to cater to the desire for user-created content and the long tail economics (including mobile operators, telecommunication service providers, traditional media publishers and search engines)
  - Greater availability of money related to advertising and new business models to monetise content
- *Institutional and Legal Drivers*
  - Rise of schemes which provide more flexible access to creative works and the right to create derivative works (eg: flexible licensing and copyright schemes such as the Creative Commons licence)
  - Rise of end-user licensing agreements which grant copyright to users for their content

Though not explicitly included in a taxonomy in their paper, Wunsch-Vincent & Vickery [2007] then describe various aspects of UCC/UGC, and from the section headings I have extracted the following taxonomy, as well. These would cut across the taxonomy of drivers, presented above.

- *Types of UCC*: Text, novel and poetry; photos/images; music and audio; video and film; citizen journalism; educational content; mobile content; and virtual content.
- *Distribution platforms*: Blogs; wikis and other text-based collaboration formats; sites allowing feedback on written works; group-based aggregation; podcasting; social network sites; virtual worlds; and content or filesharing sites.
- *Monetisation of user-created content and new business models*: Voluntary donations; charging viewers for services (pay-per-item or subscription); advertising-based models; licensing of content and technology to third parties; and selling goods and services to community.

## B. Published taxonomies of user generated content

---

- *Economic incentives along the value chain*: consumer electronics and ICT goods; software producers; ISPs and Web portals; UCC platforms and sites; users and creators; traditional media; professional content creators; search engines; Web services that profit from UCC; advertising; and marketing and brands.
- *Social impacts of user-created content*: Changed information production leading to increased user autonomy, participation and communication; cultural impacts; citizenship engagement and politics; educational and informative impact; impact on ICT and other skills; and social and legal challenges of user-created content.
- *Digital content policies*: Enhancing R&D, innovation and technology in content, networks, software and new technologies; developing a competitive, non-discriminatory framework environment (i.e. value chain and business model issues); enhancing the infrastructure (e.g. technology for digital content delivery, standards and interoperability); business and regulatory environments that balance the interests of suppliers and users, in areas such as the protection of intellectual property rights and digital rights management, without disadvantaging innovative e-business models; governments as producers and users of content (e.g. commercial re-use of public sector information); and conceptualisation, classification and measurement issues. Within the “business and regulatory environments” item, they provide a detailed classification concerning *intellectual property rights and user-created content*:
  - copyrights in the context of user-created content (Original works created by users; derivative works; and facilitating UCC creation);
  - copyrights and the terms of services of UCC sites;
  - copyrights and the liability of UCC platforms;
  - digital rights management;
  - freedom of expression;
  - information and content quality;
  - mature, inappropriate, and illegal content<sup>1</sup>;
  - safety on the Internet and awareness raising;
  - privacy and identity theft;
  - impacts of intensive Internet use;
  - network security and spam;
  - virtual worlds, property rights and taxation;
  - governments as producers and users of content; and
  - conceptualisation, classification and measurement.

---

<sup>1</sup>Of course, lumping “mature” with “inappropriate” and “illegal” makes a judgement about “mature” content that is indefensible!

## B. Published taxonomies of user generated content

- *Participative Web technologies*: Tagging; group rating and aggregation; syndication and aggregation of data; application mash-ups and open APIs; and files sharing networks.

### B.3 Gervais' taxonomy for copyright issues

They are discussed above in Section 8.4.3.

Gervais [2009] drew on the OECD taxonomy of user-created content, identifying it as a matrix of type of content against distribution platform. However, as this did not meet his needs for understanding the copyright issues, he developed his own taxonomy of user-generated content for this purpose. While this taxonomy is obviously limited, it adds an important dimension:

- *User-authored content*: Content authored without copying, derivation or adaption, and hence easy to deal with from the copyright perspective, as "the author is free to copy, upload, perform and/or make available" their content on any basis. A complication could arise when the author uses a Web site that takes a licence for the site's owner to use the content, as a condition of using the Web site. I would suggest that this usage by the site's owner no longer constitutes "user-authored content" but rather becomes "user-copied content".
- *User-derived content*: Considered by Gervais to be the most complicated category, because of the nature of the underlying right and whether or not the derivation and/or reproduction constituted *fair use* (which is determined by the use value gained by the user and the exchange value lost by the rights holder). Examples of fair use for user-derived content include critiques and parodies.
- *User-copied content*: Merely copying pre-existing content is *prima facie* infringement and hence generally illegal and illegitimate. However, it could be considered to be fair use if only a "short excerpt" is used (determined qualitatively more so than just quantitatively) or if the use is "transformative". A complication here is that the First Amendment [United States of America 1791] has been used in the USA as a defence, which is not necessarily applicable in other countries<sup>2</sup>. Examples of fair use for user-copied content are framing (including another Web site unaltered within a frame on one's own Web site, without actually copying the content of the other Web site) and "thumbnail" images of Web pages for linking to them.
- *Peer-to-peer as UGC*: The key difference between this category and "user-copied content" would appear to be that "user-copied content" should be "transformative", that is, that it does not "merely supersede the objects of the original creation" [US Court of Appeals for the Ninth Circuit 2007]. Gervais [2009] feels that while unauthorized peer-to-peer (P2P) file-sharing is generally illegal, it is not going away and controlled monetizing is the best outcome for both authors and users.

<sup>2</sup>Gervais is based in the USA and wrote from that perspective, though he was educated in Canada and formerly worked for both WIPO and WTO.

## *B. Published taxonomies of user generated content*

---

### **B.4 Budhathoki, Nedovic-Budic and Bruce's framework for VGI**

Budhathoki *et al* [2009] presented an overall framework for conceptualizing volunteered geographical information, which is discussed above in Section 8.4.4:

- *Context of Participation*
  - Personal
  - Social
  - Technological
- *Motivation*
  - Unique ethos
  - Learning
  - Career
  - Personal enrichment
  - Self-actualization
  - Self-expression
  - Self-image
  - Self-gratification/Fun
  - Re-creation
  - Social
  - Group accomplishment
  - Group attraction
  - Group maintenance
  - Identity
  - Reputation
  - Monetary
  - Instrumentality
  - Cognitive capital/self-efficacy
  - Reciprocity
  - Sense of community
  - Meeting own need
  - Freedom and creativity

## *B. Published taxonomies of user generated content*

---

- Altruism
- Trust in the underlying infrastructure
- Protective
- Structural capital
- Self-presentation
- Relation management
- Socio-political motives
- *Contribution Mechanisms*
  - Structure
  - Process
  - Norms
- *Contribution*
- *Issues*
  - Reliability
  - Quality
  - Value
  - Privacy
  - Copyright
  - Coverage
  - Credibility
  - Sustainability
  - Social justice

However, as this was a conference presentation, they did not define any of these terms, though they did provide a detailed expansion of their class *Motivation*, providing 29 motivational factors, with conceptual definitions and literature sources for them [Budhathoki *et al* 2009]. There appears to be overlaps between many of the motivational factors they list, particularly those related to self actualization, but it is not clear from their presentation material if they were merely documenting the motivational factors they had found in the literature, or if they were making value judgments on them.

Budhathoki *et al* [2009] also presented an analysis of the talk pages of OpenStreetMap [2016], which effectively gives a taxonomy of the motivations of contributors to OpenStreetMap:

- *Fulfillment of self-need*

## *B. Published taxonomies of user generated content*

---

- *Anti-corporate sentiment (unique ethos)*
- *Expectation of reciprocity*
- *Visual power of map (self-gratification)*
- *Outdoor activity (re-creation)*
- *Pride of local knowledge*
- *Concerns for a substantive issue (need)*
- *Other — explored: monetary, hobby, learning.*

### **B.5 Coleman, Georgiadou and Labonte’s nature and motivation of producers**

Coleman *et al* [2009] considered the nature and motivation of *producers* of volunteered geographical information. They characterized the contributors of VGI as seen by the early commentators into five overlapping categories, which are discussed above in Section 8.4.5::

- *Neophyte*: someone with no formal background in the subject, but with the interest, time and willingness to offer an opinion (or data).
- *Interested amateur*: someone gaining knowledge and expertise in the subject, though reading, experimenting and consulting with other colleagues and experts.
- *Expert amateur*: someone knowing much about a subject and practicing it passionately on occasion, but not relying on it for a living. Presumably, this also includes those with detailed and relevant local knowledge about their environment, as opposed to knowledge about a discipline.
- *Expert professional*: someone with the education and professional recognition in the subject to be able to rely on it for a living, and may be sued if they fail their customers.
- *Expert authority*: someone with greater knowledge and experience of the subject than the expert professional, with an established track record and in a position to lose that reputation and even their livelihood if their credibility is lost, even temporarily.

Coleman *et al* [2009] then provide four overlapping contexts in which individuals contribute VGI:

- *Mapping and navigation*
- *Social networks*
- *Civic/governmental*
- *Emergency reporting*

---

*B. Published taxonomies of user generated content*

---

They can be characterized by:

- Their humanity
- Frequency, type and degree of contributions
- Quality and veracity
- Reputation for reliability

The following are reasons to make constructive contributions:

- Altruism
- Professional or personal interest
- Intellectual stimulation
- Protection or enhancement of a personal investment
- Social reward
- Enhanced personal reputation
- Outlet for creative and independent self-expression
- Pride of place

The following are reasons to make negative contributions:

- Mischief
- Agenda
- Malice and/or criminal intent

The following types of contributions to Wikipedia could also be made to VGI repositories:

- *Constructive:*
  - Legitimate new content
  - Constructive amendments
  - Validation and repair
  - Minor edits and format changes
- *Damaging:*
  - Mass deletes
  - Nonsense
  - Spam
  - Partial deletes
  - Offensive content
  - Misinformation

## *B. Published taxonomies of user generated content*

---

### **B.6 Castelein, Grus, Cromptoets and Bregt's characterization of repositories of VGI**

Castelein *et al* [2010] characterized repositories of VGI (though their text implies they characterized VGI *per se*) from the perspective of SDI components, using the conceptual model of Rajabifard *et al* [2002], which has five core components. To this, they added thirteen characteristics to describe VGI, chosen because of ease of measurement by Web survey, objective character and clear presentation of the five SDI components. This taxonomy is good for characterization (their intended purpose) but not for our purposes, as discussed above in Section 8.4.6.

- *Policy*
  1. Whether or not registration is required to contribute.
- *Access network*
  3. If application programming interface(s) are available.
  4. Available services: download and/or upload of data.
- *Standards*
  5. If there are standard feature types and/or standard data formats for uploading.
- *Data*
  6. Total number of contributions uploaded.
  7. Data types that can be uploaded: point, line and/or polygon data.
  8. If the last update or contribution to the Web site within the last hour.
  9. If there is a thematic focus or user community with a specific theme.
  10. Geographic extent of the data: global, continent, region, etc.
  11. If the site only has VGI, or if it is combined with official data.
- *People*
  12. Number of registered users.
  13. Number of unique visitors per day.
  14. Number of unique sites linking to the site.

\*\*\*\*

---

*B. Published taxonomies of user generated content*

# Bibliography

*Please note that this thesis has been written from 2009 to 2016 and includes many references with URLs. Unfortunately, some of these links are now broken, and it has not been possible to verify them all again.*

6DISS (31 Dec 2005) *Report on the workshop and status of Internet connectivity in Southern Africa*. Deliverable report D04, 6DISS: IPv6 Dissemination and Exploitation.

**Abler**, R F (1987) *The National Science Foundation National Center for Geographic Information and Analysis*. International Journal of Geographical Information Systems, **1**(4):pp 303–326.

**Acharya**, Anurag, **Verstak**, Alex, **Suzuki**, Helder, **Henderson**, Sean, **Iakhiaev**, Mikhail, **Lin**, Cliff Chiung Yu & **Shetty**, Namit (9 Oct 2014) *Rise of the rest: The growing impact of non-elite journals*. arXiv, (1410.2217v1).

**Achenbach**, Joel (15 Apr 2013) *Scientist tries to unravel mystery of Coral Sea's ghostly island*. ADNcom. URL <http://www.adn.com/article/20130415/scientist-tries-unravel-mystery-coral-seas-ghostly-island>.

**Adams**, Benjamin & **Gahegan**, Mark (7–9 Apr 2014) *Emerging data challenges for next-generation spatial data infrastructure*. In: **Winter**, S & **Rizos**, C (eds), *Research@Locate'14*, Canberra, Australia. URL <http://ceur-ws.org/>.

**Adams**, Katherine (Jul/Aug 2002) *The Semantic Web: Differentiating between taxonomies and ontologies*. Online, **26**(4):pp 20–23.

**Addley**, Esther (24 Sep 2014) *Threat to post naked photographs of Emma Watson appears to be hoax*. The Guardian. URL <http://www.theguardian.com/film/2014/sep/24/threat-post-naked-photographs-emma-watson-hoax-4chan>.

**Adeleye**, Omokhoa A & **Adebamowo**, Clement A (Dec 2012) *Factors associated with research wrongdoing in Nigeria*. Journal of Empirical Research on Human Research Ethics, **7**(5):pp 15–24, doi: 10.1525/jer.2012.7.5.15.

**Adie**, Euan, **Adler**, Elizabeth M, **Ahmad**, Sharon, **Albertine**, Kurt H, **Alberts**, Bruce &

## Bibliography

- 150 others** (16 Dec 2012) *San Francisco Declaration on Research Assessment: Putting science into the assessment of research*. San Francisco, CA, USA.
- AFP Agency Staff** (4 Dec 2014) *YouTube counter can't handle Gangnam Style*. Business Day Live. URL <http://www.bdlive.co.za/life/gadgets/2014/12/04/youtube-counter-cant-handle-gangnam-style>.
- AFP Agency Staff** (5 Oct 2015a) *Embattled Uber faces global crackdown*. Fin24com. URL <http://www.fin24.com/Tech/News/Embattled-Uber-faces-global-crackdown-20151005>.
- AFP Agency Staff** (1 Aug 2015b) *MtGox bitcoin exchange chief executive Mark Karpeles arrested in Japan*. Australian Broadcasting Corporation. URL <http://www.abc.net.au/news/2015-08-01/mtgox-bitcoin-exchange-ceo-mark-karpeles-arrested-in-japan/6665836>.
- AFP Agency Staff** (27 Sep 2015c) *Uber, Airbnb a tax threat for economies*. Fin24com. URL <http://www.fin24.com/Economy/Uber-Airbnb-a-tax-threat-for-economies-20150927>.
- AFP Agency Staff** (1 Jan 2016) *Uber hits one billion rides*. Fin24com. URL <http://www.fin24.com/Tech/News/uber-hits-one-billion-rides-20160101>.
- AFP, SAPA-DPA** (12 Sep 2013) *US spying: Yahoo! chief fears jail for going against NSA demands*. Mail & Guardian Online. URL <http://mg.co.za/print/2013-09-12-us-spying-yahoo-chief-feared-jail-for-going-against-nsa-demands>.
- Africa News Agency** (26 Jul 2015) *SA's Chad Ho wins gold at Fina world champs*. eNCA. URL <http://www.enca.com/sport/sas-chad-ho-wins-gold-fina-world-champs>.
- Ái, Võ Văn, Faulkner, Penelope & Ye, Shiwei** (Feb 2013) *Bloggers and netizens behind bars: Restrictions on internet freedom in vietnam*. Tech rep, International Federation for Human Rights (FIDH) and the Vietnam Committee on Human Rights of Que Me: Action for Democracy in Vietnam.
- Albeanu, Catalina** (28 Oct 2014) *'space journalism': How ProPublica tells stories using satellites*. Journalismcoulk. URL <http://www.journalism.co.uk/news/space-journalism-how-propublica-tells-stories-using-satellite-imagery/s2/a562960>.
- Alexa** (2014) *Alexa — Actionable Analytics for the Web. Home page*. Alexa Internet, Inc. URL <http://www.alexa.com/>.
- Alexander, James** (1736) *A brief narrative of the case and tryal of John Peter Zenger, printer of the New-York Weekly Journal*. New-York Weekly Journal. Reproduced by the Historical Society of the Courts of the State of New York, URL <http://www.courts.state.ny.us/history/zenger.htm>.
- Alexander, Ruth** (20 Feb 2012) *Are search engine result figures accurate?* BBC News Magazine. URL <http://www.bbc.co.uk/news/magazine-17068044>.

## Bibliography

---

- Alexander**, Ruth (19 Apr 2013) *Reinhart, Rogoff...and Herndon: The student who caught out the profs*. BBC News Magazine. URL <http://www.bbc.co.uk/news/magazine-22223190>.
- Alfred**, Charlotte (16 Oct 2015) *The citizen journalists challenging Assad And Putin's story of war*. The World Post. URL [http://www.huffingtonpost.com/entry/eliot-higgins-putin-syria\\_us\\_561fc877e4b028dd7ea6e15c](http://www.huffingtonpost.com/entry/eliot-higgins-putin-syria_us_561fc877e4b028dd7ea6e15c).
- Alfreds**, Gareth, Duncanand van Zyl (23 Oct 2015b) *Last throw of the dice for Mxit*. Fin24. URL <http://www.fin24.com/Tech/News/Lastthrow-of-the-dice-for-Mxit-20151023>.
- All Things Spatial** (13 Oct 2009) *Ed parsons on spatial data infrastructure*. All Things Statial. URL <http://all-things-spatial.blogspot.com/2009/10/ed-parsons-on-spatial-data>.
- Allen**, Michael A (2010) *On the current obsession with publication statistics*. ScienceAsia, 36:pp 1–5.
- Allen**, Naomi, **Kreps**, Jason, **Osowski**, Kaydi, **Casillas**, Andrea, **Lemmo**, Thomas, **Wong**, Christopher, **Deveau-Rosen**, Jason, **Merante**, Joseph, **Webbink**, Mark & **Murphy**, Michael (Jun 2009) *Peer to Patent: Second anniversary report*. Tech rep, Center for Patent Innovations, The Institute for Information Law and Policy, New York Law School, 57 Worth Street, New York, NY 10013-2960. URL [http://dotank.nyls.edu/communitypatent/CPI\\_P2P\\_YearTwo\\_lo.pdf](http://dotank.nyls.edu/communitypatent/CPI_P2P_YearTwo_lo.pdf).
- Allen**, Naomi, **Merante**, Joseph, **Tham**, Yeen, **Ingham**, Joanne, **Noveck**, Beth Simone, **Webbink**, Mark, **Johnson**, Bridgette, **Stock**, William & **Wong**, Christopher (Jun 2008) *Peer to Patent: First anniversary report*. Tech rep, Center for Patent Innovations, The Institute for Information Law and Policy, New York Law School, 57 Worth Street, New York, NY 10013-2960. URL <http://dotank.nyls.edu/communitypatent/P2Panniversaryreport.pdf>.
- Altman**, Alex (Sep 2014) *Why terrorists love Twitter*. Time. URL <http://time.com/3319278/isis-isil-twitter/>.
- Ambrose**, Steven (18 Jun 2009) Quoted in the column “TheInsider” in Business Day, page 8.
- Ana**, Joseph, **Koehlmoos**, Tracey, **Smith**, Richard & **Yan**, Lijing L (2013) *Research misconduct in low- and middle-income countries*. PLoS Medicine, 10(3):p e1001315, doi: 10.1371/journal.pmed.1001315.
- Ananthaswamy**, Anil (20 Jul 2011) *Welcome to the age of the splinternet*. New Scientist, (2821).
- Anderson**, Chris (Oct 2004) *The long tail*. Wired, (12.10). URL [http://www.wired.com/wired/archive/12.10/tail\\_pr.html](http://www.wired.com/wired/archive/12.10/tail_pr.html).
- Anonymous** (2011) *GSIM. Information Systems in Official Statistics: What everyone should know about statistical metadata*. Accessed 4 July 2014, URL <http://isosmeta.wordpress.com/>.

- ANSI (1986) *INCITS 4-1986[R2012], Information systems — Coded Character Sets — 7-Bit American National Standard Code for Information Interchange (7-bit ASCII)*. American National Standards Institute (ANSI), Washington, DC, USA. Previously known as ANSI X3.4-1986.
- Anthony**, Sebastian (2 Apr 2012) *Internet-induced fear culture (or: Why Girls Around Me isn't the problem)*. Extremetech. URL <http://www.extremetech.com/extreme/124531-internet-induced-fear-culture-or-why-girls-around-me-isnt-the-problem>.
- APPSI (17 Feb 2010) *Minutes of the meeting of 17 February 2010*. Advisory Panel on Public Sector Information.
- Armenakis**, C., **Dow**, A., **Gray**, L., **Mattar**, K. & **Van Der Kooij**, M. (1729) *Evaluation of spaceborne InSAR generated DEM*. *Orbit*, **24242**(4569):p 2072. *This reference has been included verbatim from the BibTeX version on Google Scholar, as obtained on 7 July 2009. The year of this publication, its volume and its number are patently wrong, so it is not certain what, if anything, in the reference is correct!*
- Armstrong**, J Scott (Jun 1982) *Barriers to scientific contributions: The author's formula*. *Behavioral and Brain Sciences*, **5**:pp 197–199.
- Armstrong**, J Scott & **Overton**, Terry S (1977) *Estimating nonresponse bias in mail surveys*. *Journal of Marketing Research*, **14**(3):pp 396–402.
- Arnold**, Doug (5 Feb 2012) *More reasons to support the Elsevier boycott*. International Mathematical Union Blog. URL [http://blog.mathunion.org/journals/?tx\\_t3blog\\_pi1\[blogList\]\[showUid\]=30](http://blog.mathunion.org/journals/?tx_t3blog_pi1[blogList][showUid]=30).
- Arnold**, Lesley M & **Wright**, Graeme L (2005) *Analysing product-specific behaviour to support process dependent updates in a dynamic spatial updating model*. *Transactions in GIS*, **9**(3):pp 397–419.
- Asheim**, Lester (Sep 1953) *Selection and censorship: A reappraisal*. *Wilson Library Bulletin* (R), **58** N:pp 180–184.
- Asheim**, Lester (Nov 1983) *Not censorship but selection*. *Wilson Library Bulletin*, **28**:pp 63–67. Made available by the American Library Association, URL <http://www.ala.org/advocacy/intfreedom/censorshipfirstamendmentissues/notcensorship>.
- Ashley**, Kevin (10 Aug 2013) *Data quality and curation*. *Data Science Journal*, **12**:pp GRDI65–GRDI68.
- Ashton**, Kevin (17 Apr 2013a) *How to become internet famous for \$68*. Quartz: Tweeto Ergo Sum. URL <http://qz.com/74937/how-to-become-internet-famous-without-ever-existing/>.
- Ashton**, Kevin (28 Mar 2013b) *You didn't make the harlem shake go viral — corporations did*. Quartz: Tweeto Ergo Sum. URL <http://qz.com/67991/you-didnt-make-the-harlem-shake-go-viral-corporations-did/>.
- Assange**, Julian (22 Jun 2013b) *Statement by julian assange after one year in ecuadorian*

## Bibliography

---

- embassy*. Wikileaks. URL <http://wikileaks.org/Statement-by-Julian-Assange-after,249.html>.
- Ather**, Aamer (2009) *A quality analysis of OpenStreetMap data*. MEng, University College London, London, United Kingdom.
- Atkinson**, Malcolm, **Bancilhon**, Francois, **DeWitt**, David, **Dittrich**, Klaus, **Maier**, David & **Zdonik**, Stanley (Dec 1989) *The object-oriented database system manifesto*. In: *Proceedings of the First International Conference on Deductive and Object-Oriented Databases*, Kyoto, Japan, pp 223–40. URL <http://www.cs.cmu.edu/afs/cs.cmu.edu/user/clamen/OODBMS/Manifesto/>.
- Attwell**, Arthur (21 May 2013) *Why i publish ebooks on paper for South Africans*. Publishing Perspectives. URL <http://publishingperspectives.com/2013/05/why-i-publish-ebooks-on-paper-for-south-africans/>.
- Austin**, Dan (2012) *Wiping out spam*. Map Makerpedia. URL <https://sites.google.com/site/mapmakerpedia/maps-101/wiping-out-spam>.
- Austin**, Dan (27 Nov 2013) *Google MapMaker update summary: One database to rule them all*. Blumenthals. URL <http://blumenthals.com/blog/2013/11/27/google-mapmaker-update-summary-one-database-to-rule-them-all/>.
- Austin**, Dan (18 Mar 2014) *Google Maps report a problem: Does it work for local spam?* Blumenthals. URL <http://blumenthals.com/blog/2014/03/18/google-maps-report-a-problem-does-it-work-for-local-spam/>.
- AutoCarto (29 Mar–3 Apr 1987) *Auto-Carto VIII: Proceedings of the International Symposium on Computer-Assisted Cartography*, Baltimore, MD, USA. URL <http://www.mapcontext.com/autocarto/proceedings/auto-carto-8/>.
- Azzaman** (26 Dec 2012) *Arab bloggers still face harassment, persecution*. Al-Monitor: The Pulse of the Middle East. Translated by Rani Geha, URL <http://www.al-monitor.com/pulse/culture/2012/12/blogging-in-the-arab-world>.
- Baker**, Paul & **Potts**, Amanda (2013) ‘*Why do white people have thin lips?*’ Google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies*, 10(2):pp 187–204, doi: 10.1080/17405904.2012.744320.
- Bakshy**, Eytan, **Messing**, Solomon<sup>1</sup> & **Adamic**, Lada (7 May 2015) *Exposure to ideologically diverse news and opinion on Facebook*. *Science*, doi: 10.1126/science.aaa1160.
- Balkin**, Jack M (5 Mar 2014) *Information fiduciaries in the digital age*. Balkinization. URL <http://balkin.blogspot.com/2014/03/information-fiduciaries-in-digital-age>.
- Ball**, James, **Borger**, Julian & **Greenwald**, Glenn (6 Sep 2013a) *Revealed: how US and UK spy agencies defeat internet privacy and security*. *Guardian Weekly*. URL <http://www.theguardian.com/world/2013/sep/05/nsa-gchq-encryption-codes-security>.
- Ball**, James, **Schneier**, Bruce & **Greenwald**, Glenn (4 Oct 2013b) *NSA and GCHQ tar-*

## Bibliography

- get Tor network that protects anonymity of web users*. The Guardian. URL <http://www.theguardian.com/world/2013/oct/04/nsa-gchq-attack-tor-network-encryption>.
- Ball**, Matt (31 Jan 2010a) *What can be learned from the volunteer mapping efforts for Haiti?* Spatial Sustain. URL <http://www.sensysmag.com/spatialsustain/what-can-be-learned-from-the-volunteer-mapping-efforts-for-haiti>.
- Ball**, Matt (3 Jun 2013) *Must map monetization mean guiding rather than enabling exploration?* Sensors & Systems: Making sense of global change. URL <http://www.sensorsandsystems.com/dialog/perspectives/30574-must-map-monetization-mean-guiding-rather-than-enabling-exploration>.
- Barber**, Nicholas (4 Aug 2015) *Comedy in the age of outrage: When jokes go too far*. BBC Culture. URL <http://www.bbc.com/culture/story/20150804-comedy-in-the-age-of-outrage-when-jokes-go-too-far>.
- Barford**, Vanessa (17 Aug 2011) *England riots: What are the Post-it note 'love walls' all about?* BBC News. URL <http://www.bbc.co.uk/news/magazine-14548710>.
- Barry**, Ellen & **Raj**, Suhasini (8 Dec 2014) *Uber is banned in Delhi after driver is accused of rape*. The New York Times. URL <http://nyti.ms/1zEJ3LE>.
- Basiouka**, Sofia & **Potsiou**, Chryssy (22 Aug 2013) *The volunteered geographic information in cadastre: perspectives and citizens' motivations over potential participation in mapping*. GeoJournal, doi: 10.1007/s10708-013-9497-7.
- Bassett**, Caroline (4 Mar 2013) *Silence, delirium, lies?* First Monday, **18**(3), doi: 10.5210/fm.v18i3.4617. URL <http://firstmonday.org/ojs/index.php/fm/article/view/4617/3420>.
- Batty**, Michael, **Hudson-Smith**, Andrew, **Milton**, Richard & **Crooks**, Andrew (2010) *Map mashups, Web 2.0 and the GIS revolution*. Annals of GIS, **16**(1):pp 1–13.
- BBC** (29 Apr 2011a) *Amazon apologises for cloud fault one week on*. BBC News: Business. URL <http://www.bbc.co.uk/news/business-13242782>.
- BBC** (21 Apr 2011b) *Jihadists use mobiles as propaganda tools*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-13149571>.
- BBC** (9 Aug 2011d) *Secret net Tor asks users to sign up to cloud services*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-15834476>.
- BBC** (8 Sep 2012a) *Author Roth rebukes Wikipedia over Human Stain edit*. BBC News: Entertainment & Arts. URL <http://www.bbc.co.uk/news/entertainment-arts-19527797>.
- BBC** (28 Nov 2012b) *China paper carries Onion Kim Jong-un 'heart-throb' spoof*. BBC News: Asia. URL <http://www.bbc.co.uk/news/world-asia-20518929>.
- BBC** (22 Jul 2012c) *Chirp app sends smartphone data via 'digital birdsong'*. BBC News: Technology. URL <http://www.bbc.com/news/technology-18927928>.
- BBC** (17 Feb 2012d) *Google cookies 'bypassed Safari privacy protection'*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-17076670>.

## Bibliography

---

- BBC** (15 Feb 2012e) *Social apps 'harvest smartphone contacts'*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-17051910>.
- BBC** (23 Jan 2012f) *Storage sites unnerved by Megaupload action*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-16679174>.
- BBC** (9 Aug 2013a) *Confused xerox copiers rewrite documents, expert finds*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-23588202>.
- BBC** (29 Jan 2013b) *Google expands North Korea map coverage*. BBC News: Business. URL <http://www.bbc.co.uk/news/business-21226623>.
- BBC** (15 Feb 2013c) *Meteor highlights rise of dashboard cameras in Russia*. BBC News: Europe. URL <http://www.bbc.co.uk/news/world-europe-21478361>.
- BBC** (30 Jun 2013d) *Saudi Arabia court jails seven Facebook cyber activists*. BBC News: Middle East. URL <http://www.bbc.co.uk/news/world-middle-east-23119656>.
- BBC** (30 Jul 2013e) *Thousands of abusive electronic message cases reach court*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-23502291>.
- BBC** (27 Sep 2013f) *Yelp admits a quarter of submitted reviews are fake*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-24299742>.
- BBC** (10 Nov 2014a) *L'Aquila quake: Scientists see convictions overturned*. BBC News: Europe. URL <http://www.bbc.com/news/world-europe-29996872>.
- BBC** (29 Sep 2014b) *Peter Nunn jailed for Twitter abuse of MP Stella Creasy*. BBC News: England. URL <http://www.bbc.com/news/uk-england-29411031>.
- BBC** (31 Mar 2014c) *Turkey hijacks servers in social media crackdown*. BBC News: Technology. URL <http://www.bbc.com/news/technology-26818104>.
- BBC** (7 Aug 2015a) *Bangladesh blogger Niloy Neel hacked to death in Dhaka*. BBC News. URL <http://www.bbc.com/news/world-asia-33819032>.
- BBC** (9 Jan 2015b) *Saudi blogger Badawi 'flogged for Islam insult'*. BBC News: Middle East. URL <http://www.bbc.com/news/world-middle-east-30744693>.
- BBC** (8 Jan 2016) *Drug lab raid ruins Australian family's Airbnb holiday*. BBC News. URL <http://www.bbc.com/news/world-australia-35258683>.
- Beall, Jeffrey** (2006a) *The death of metadata*. The Serial Librarian: From the Printed Page to the Digital Age, **51**(2).
- Beall, Jeffrey** (Feb 2006b) *Metadata and data quality problems in the digital library*. Journal of Digital Information, **6**(3).
- Beall, Jeffrey** (2010) *How Google uses metadata to improve search results*. The Serial Librarian: From the Printed Page to the Digital Age, **59**(1).
- Beall, Jeffrey** (13 Sep 2012) *Predatory publishers are corrupting open access*. Nature, **489**:p 179.

- Beall, Jeffrey** (16 Sep 2014a) *Bogus “center” provides quick, easy, and cheap publishing*. Scholarly Open Access: Critical analysis of scholarly open-access publishing. URL <http://scholarlyoa.com/2014/09/16/bogus-center-provides-quick-easy-and-cheap-publishing/>.
- Beall, Jeffrey** (4 Nov 2014c) *Google Scholar is filled with junk science*. Scholarly Open Access: Critical analysis of scholarly open-access publishing. URL [scholarlyoa.com/2014/11/04/google-scholar-is-filled-with-junk-science/](http://scholarlyoa.com/2014/11/04/google-scholar-is-filled-with-junk-science/).
- Beall, Jeffrey** (2 Jan 2014d) *List of predatory publishers 2014*. Scholarly Open Access: Critical analysis of scholarly open-access publishing. URL <http://scholarlyoa.com/2014/01/02/list-of-predatory-publishers-2014/>.
- Beall, Jeffrey** (29 Jul 2014e) *Meta-analyses and the problems of duplicate publication and plagiarism*. Scholarly Open Access: Critical analysis of scholarly open-access publishing. URL <http://scholarlyoa.com/?s=Meta-analyses+and+the+Problems+of+Duplicate+Publication+and+Plagiarism>.
- Beall, Jeffrey** (Feb 2014g) *Unintended consequences: The rise of predatory publishers and the future of scholarly publishing*. Editorial Office News:pp 4–6. URL <http://www.ismte.org/>.
- Beall, Jeffrey** (24 Jan 2014h) *University of Pristina Rector under fire for publishing in predatory journals*. Scholarly Open Access: Critical analysis of scholarly open-access publishing. URL <http://scholarlyoa.com/2014/01/24/university-of-pristina-rector-under-fire-for-publishing-in-predatory-journals/>.
- Beall, Jeffrey** (1 Jan 2015) *Criteria for determining predatory open-access publishers*. Scholarly Open Access: Critical analysis of scholarly open-access publishing. URL <http://scholarlyoa.com>.
- Beck, G, Bethe, H & Riezler, W** (9 Jan 1931) *Bemerkung zur Quantentheorie der Nullpunktstemperatur*. Die Naturwissenschaften, 2:p 39, doi: 10.1007/BF01523870. Translation available at: URL <http://www.math.tohoku.ac.jp/~kuroki/Sokal/misc/bethespoof.html>.
- Beckett, Andy** (26 Nov 2009) *The dark side of the internet*. The Guardian. URL <http://www.theguardian.com/technology/2009/nov/26/dark-side-internet-freenet>.
- Bédard, Yvan** (18 May 2012) *Geospatial data quality awareness, the next challenge: Are we ready?* In: *International Workshop on Geospatial Data Quality Legal, ethical and technical aspects*, Québec City, Canada. URL <http://dataquality.scg.ulaval.ca>.
- Béjar, Rubén, Latre, Miguel Á, Nogueras-Iso, Javier, Muro-Medrano, Pedro R & Zarazaga-Soria, F Javier** (2011) *An RM-ODP enterprise view for spatial data infrastructures*. Computer Standards & Interfaces, doi: 10.1016/j.csi.2011.10.001. In press.
- Ben Yahia, Sadok & Mephu Nguifo, Engelbert** (eds) (30 Oct–1 Nov 2006) *4th International Conference on Concept Lattices and Their Applications*, Université Centrale, Tunis, Yasmine Hammamet, Tunisia.

## Bibliography

- Bentham**, Jeremy (1787) *Panopticon; or the inspection-house: Containing the idea of a new principle of construction applicable to any sort of establishment, in which persons of any description are to be kept under inspection; and in particular to penitentiary-houses, prisons, houses of industry, work-houses, poor-houses, lazarettos, manufactories, hospitals, mad-houses, and schools: With a plan of management adapted to the principle*. Letters of Bentham. Transcription and HTML by Cartome, 16 June 2001, from Bentham, Jeremy *The Panopticon Writings*. Ed. Miran Bozovic (London: Verso, 1995). p. 29-95, URL <http://cartome.org/panopticon2.htm>.
- Bergman**, Michael K (Aug 2001) *White Paper: The Deep Web: Surfacing Hidden Value*. Journal of Electronic Publishing, 7(1). URL <http://dx.doi.org/10.3998/3336451.0007.104>.
- Berners-Lee**, Tim (18 Jun 2009) *Linked data*. World Wide Web Consortium. First published 2006-07-27, updated 2010, URL <http://www.w3.org/DesignIssues/LinkedData.html>.
- Berners-Lee**, Tim (nd) *Answers for young people*. World Wide Web Consortium (W3C). Accessed 2015-12-30, URL <https://www.w3.org/People/Berners-Lee/Kids.html>.
- Berners-Lee**, Tim, **Hendler**, James & **Lassila**, Ora (May 2001) *The Semantic Web*. Scientific American, 284(5):pp 28–37. URL <http://www.scientificamerican.com/article.cfm?id=the-semantic-web>.
- Bessero**, Gilles, **Brodeur**, Jean, **Coetzee**, Serena, **Østensen**, Olaf, **Pharaoh**, Anthony & **Reed**, Carl (6 Jun 2013) *The UN-GGIM inventory of issues and geographic information standardization*. Tech rep, ISO/TC 211, *Geographic information/Geomatics*, in cooperation with the Open Geospatial Consortium (OGC) and the International Hydrographic Organization (IHO).
- Bevel**, Tom & **Gardner**, Ross M (2008) *Bloodstain Pattern Analysis: With an Introduction to Crime Scene Reconstruction*. CRC Press, 3rd edn, ISBN 978-1-4200-5268-8, 440 pp.
- Biagio**, Dina (Aug 2013) *More patents than the space shuttle*. EngineerIT. URL <http://www.ee.co.za/article/spoor-and-fishers-100-04-more-patents-than-the-space-shuttle.html>.
- Bianchini**, Laurence (19 Jul 2011) *Scientific publication: the model and scandals*. MyScienceWork. A version of this article in French exists: *L'édition scientifique: son modèle, ses scandales*, URL <https://www.mysciencework.com/omniscience/scientific-publication-the-model-and-scandals>.
- Bidwell**, Nicola J, **Reitmaier**, Thomas, **Rey-Moreno**, Carlos, **Roro**, Zukile, **Siya**, Masbulele Jay & **Dlutu**, Bongiwe (May 2013) *Timely relations in rural Africa*. In: 12th International Conference on Social Implications of Computers in Developing Countries, Ocho Rios, Jamaica.
- Big Brother Watch** (nd) *Briefing note: Why communications data (metadata) matter. what are communications data?* Tech rep, Big Brother Watch: Defending Civil Liberties, Protecting Privacy, London, United Kingdom. URL <http://www.bigbrotherwatch.org.uk>.

- Binder**, Antje (20 Oct 2012) *Plagiarist hunters defend academic standards*. DWDE. URL <http://dw.de/p/16TCP>.
- Bizer**, Christian, **Heath**, Tom & **Berners-Lee**, Tim (2009) *Linked data — the story so far*. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3). URL <http://linkeddata.org/docs/ijswis-special-issue>.
- Blana**, Natalia & **Tsoulos**, Lysandros (23–28 Aug 2015) *Development of a methodology for map quality assessment*. In: Robbi Sluter et al [2015b].
- Bodenhamer**, David J, **Harris**, Trevor M & **Corrigan**, John (2013) *Spatial narratives and deep maps: A special report*. International Journal of Humanities and Arts Computing, 7(1–2):pp 170–175.
- Bohannon**, John (4 Oct 2013) *Who's afraid of peer review?* Science, 342:pp 60–65.
- Bohr**, Niels (Jul 1937) *Causality and complementarity*. Philosophy of Science, 4(3):pp 289–298.
- Bolstad**, P (2005) *GIS Fundamentals*. Eider Press.
- Bolton**, Kevin (2013) *Navigation — can a GPS replace the paper map*. PositionIT. URL <http://eepublishers.co.za/printarticle/gps-warehouse-284.html>.
- Bonchek**, Mark & **Cornfield**, Gene (16 Jul 2013) *The coming branded-currency revolution*. Harvard Business Review. URL [http://blogs.hbr.org/2013/07/the-coming-branded-cur2014-10-16\\_13:15:00\\_rency-re/](http://blogs.hbr.org/2013/07/the-coming-branded-cur2014-10-16_13:15:00_rency-re/).
- Bonnevie**, Bo T (2011) *Some considerations when comparing SABAP 1 with SABAP 2 data*. Ostrich: Journal of African Ornithology, 82(2):pp 161–162.
- Borba**, Rogério L R, **Strauch**, Julia C M, **Souza**, Jano M & **Coleman**, David J (23–28 Aug 2015) *Analysis of OpenStreetMap to support an official hybrid database*. In: Robbi Sluter et al [2015b].
- Borger**, Julian (24 Sep 2013) *Brazilian president: US surveillance a 'breach of international law'*. The Guardian. URL <http://www.theguardian.com/world/2013/sep/24/brazil-president-un-speech-nsa-surveillance>.
- Borges**, Jorge Luis (Mar 1946) *Del rigor en la ciencia (on exactitude in science)*. Los Anales de Buenos Aires, 1(3).
- Borges**, Jorge Luis (1952) *Otras Inquisiciones*, chap The analytical language of John Wilkins. Sur, Buenos Aires, Argentina. Translated from the Spanish 'El idioma analítico de John Wilkins' by Lilia Graciela Vázquez; edited by Jan Frederik Solem with assistance from Bjørn Are Davidsen and Rolf Andersen, URL <http://www.alamut.com/subj/artiface/language/johnWilkins.html>.
- Boulton**, Clint (17 Feb 2015) *The hidden waste and expense of cloud computing*. Business Day. Syndicated from The Wall Street Journal, URL <http://www.bdlive.co.za/life/gadgets/2015/02/17/the-hidden-waste-and-expense-of-cloud-computing>.

## Bibliography

---

- Boyd, Douglas A** (2012) *Enhancing discovery: Connecting users to your oral history collections online*. In: Boyd *et al* [2012]. URL <http://ohda.matrix.msu.edu/2012/06/enhancingdiscovery/>.
- Boyd, Douglas A, Cohen, Steve, Rakerd, Brad & Rehberger, Dean** (eds) (2012) *Oral History in the Digital Age*, Institute of Museum and Library Services, Washington, DC, USA. URL <http://ohda.matrix.msu.edu/>.
- Boyle, James** (22 Jan 2007) *Text is free, we make our money on volume(s)*. FTcom — Financial Times. URL <http://www.ft.com/cms/s/2/b46f5a58-aa2e-11db-83b0-0000779e2340>.
- Bradner, Eric** (7 Oct 2014) *Twitter wants to let users know about national security-related data requests*. CNN. URL <http://edition.cnn.com/2014/10/07/politics/twitter-sues-u-s-government/>.
- Brand, Russell** (24 Oct 2013) *Russell Brand on revolution: "we no longer have the luxury of tradition"*. New Statesman. URL <http://www.newstatesman.com/politics/2013/10/russell-brand-on-revolution>.
- Brandeis, Louis D** (4 Jun 1928) *Olmstead v. United States/Opinion of the Court*. United States Supreme Court, (277 U.S. 438). Dissenting opinion, URL <http://en.wikisource.org/w/index.php?oldid=3045965>.
- Bravo, João Vitor Meza, Camboim, Silvana Philippi, De Mendonça, André Luiz Alencar & Sluter, Claudia Robbi** (Jul–Sep 2015) *A compatibilidade dos metadados disponíveis em sistemas VGI com o perfil de metadados empregado na Infraestrutura Nacional de Dados Espaciais do Brasil (INDE-BR)*. BCG — Boletim de Ciências Geodésicas, **21**(3):pp 465–483, doi: 10.1590/S1982-21702015000300026. The compatibility of metadata available at VGI systems with the metadata profile employed at the Brazilian National Spatial Data Infrastructure (INDE-BR).
- Brembs, Björn, Button, Katherine & Munafò, Marcus** (24 Jun 2013) *Deep impact: unintended consequences of journal rank*. *Frontiers in Human Neuroscience*, **7**(291):pp 1–12.
- Bremmen, Nur** (9 Jan 2015) *Ten predictions for the immediate future of mobile*. Business Day Live, syndicated from Memeburncom. URL <http://www.bdlive.co.za/life/gadgets/2015/01/09/ten-predictions-for-the-immediate-future-of-mobile>.
- Brunon-Ernst, Anne** (2012) *Introduction*. In: *Beyond Foucault: New perspectives on Bentham's Panopticon*, Ashgate Publishing, Ltd.
- Brunon-Ernst, Anne** (20 Nov 2015) *The fallacy of informed consent: Lexical and textual markers in some user agreements of the e-commerce industry*. In: *Ecole de Droit (Sciences Po, Paris) Nudge Seminar*, Assas University. Submitted to *Revista Alicantina de Estudios Ingleses (RAEI)*.
- Buchroithner, Manfred F** (ed) (25–30 Aug 2013) *26th International Cartographic Conference (ICC 2013)*, Dresden, Germany, ISBN 978-1-907075-06-3.
- Buchroithner, Manfred F, Prechtel, Nikolas & Burghardt, Dirk** (eds) (25–30 Aug 2013) *Cartography from Pole to Pole: Selected Contributions to the XXVIth International Confer-*

- ence of the ICA, Dresden 2013. Lecture Notes in Geoinformation and Cartography: Publications of the International Cartographic Association (ICA), Springer-Verlag, ISBN 978-3-642-32617-2, doi: 10.1007/978-3-642-32618-9.
- Buckels, Erin E, Trapnell, Paul D & L, Paulhus Delroy** (2014) *Trolls just want to have fun. Personality and Individual Differences*, **67**:pp 97–102.
- Budhathoki, Nama Raj, Bruce, Bertram ‘Chip’ & Nedovic-Budic, Zorica** (2008) *Reconceptualizing the role of the user of spatial data infrastructure*. *GeoJournal*, **72**:pp 149–160.
- Budhathoki, Nama Raj, Nedovic-Budic, Zorica & Bruce, Bertram ‘Chip’** (15–19 Jun 2009) *A framework for understanding participants’ motivation in voluntary contribution of geographic information*. In: van Loenen et al [2009b].
- Budhathoki, Nama Raj, Nedovic-Budic, Zorica & Bruce, Bertram ‘Chip’** (2010) *An interdisciplinary frame for understanding volunteered geographic information*. *Geomatica*, **64**(1):pp 11–26.
- Bělohávek, Radim** (2008) *Introduction to formal concept analysis*. Tech rep, Department of Computer Science, Palacky University, Olomouc, Czech Republic. URL <http://belohlavek.inf.upol.cz/vyuka/IntroFCA.pdf>.
- Bulhak, Andrew C** (1 Apr 1996) *On the simulation of postmodernism and mental debility using recursive transition networks*. Tech Rep 96/264, Department of Computer Science, Monash University. URL [http://www.csse.monash.edu.au/cgi-bin/pub\\_search?104+1996+bulhak+Postmodernism](http://www.csse.monash.edu.au/cgi-bin/pub_search?104+1996+bulhak+Postmodernism).
- Burghart, D Brian** (Aug 2014) *What I’ve learned from two years collecting data on police killings*. Gawker. URL <http://gawker.com/what-ive-learned-from-two-years-collecting-data-on-poli-1625472836>.
- Burmeister, Peter** (3 Apr 2003) *Formal Concept Analysis with ConImp: Introduction to the Basic Features*. URL <http://www.mathematik.tu-darmstadt.de/~burmeister/ConImpIntro.pdf>.
- Burrows, John, Brett, Cate Honoré Brett, Hayward, Rachel & Hayward, Joanna** (Aug 2012) *Harmful digital communications: The adequacy of the current sanctions and remedies*. Ministerial briefing paper, Law Commission, Wellington, New Zealand.
- Bush, Randy** (1993) *FidoNet: Technology, Use, Tools, and History*. Summary report, FidoNet. URL [http://www.fidonet.org/inet92\\_Randy\\_Bush.txt](http://www.fidonet.org/inet92_Randy_Bush.txt).
- Busse, Anja** (Dec 2015) *Make an insulin pump and diabetic accessories for american girl dolls*. Changeorg. URL <https://www.change.org/p/american-girl-make-insulin-pumps-and-diabetic-accessories-for-american-girl-dolls>.
- Butenuth, Matthias, Gösseln, Guido v, Tiedge, Michael, Heipke, Christian, Lipeck, Udo & Sester, Monika** (2007) *Integration of heterogeneous geospatial data in a federated database*. *ISPRS Journal of Photogrammetry and Remote Sensing*, **62**(5):pp 328–346.
- Butler, Declan** (Feb 2006) *The web-wide world*. *Nature*, **439**(16).
- Butler, Declan** (28 Mar 2013) *The dark side of publishing*. *Nature*, **495**:pp 433–435.

## Bibliography

---

- Butterly**, Amelia (1 Sep 2014) *Jennifer Lawrence nude photos leaked 'after iCloud hack'*. BBC Newsbeat: Entertainment. URL <http://www.bbc.co.uk/newsbeat/29008876>.
- Bytyci**, Fatos (8 Feb 2014) *Kosovo's head of university quits after violent protests*. Reuters. URL <http://www.reuters.com/assets/print?aid=USBREA170T020140208>.
- Cai**, Li & **Zhu**, Yangyong (22 May 2015) *The challenges of data quality and data quality assessment in the big data era*. Data Science Journal, **14**(2):pp 1–10, doi: 10.5334/dsj-2015-002.
- Callaway**, Ewen (3 Nov 2011) *Report finds massive fraud at Dutch universities*. Nature, **479**:p 15.
- Camboim**, Silvana Philippi, **Meza Bravo**, João Vitor & **Robbi Sluter**, Claudia (2015) *An investigation into the completeness of, and the updates to, OpenStreetMap data in a heterogeneous area in Brazil*. ISPRS International Journal of Geo-Information, **4**:pp 1366–1388.
- Camponovo**, Michael E & **Freundschuh**, Scott M (2014) *Assessing uncertainty in VGI for emergency response*. Cartography and Geographic Information Science, doi: 10.1080/15230406.2014.950332.
- Caron**, Claude, **Roche**, Stéphane, **Larfouilloux**, Julien & **Hadaya**, Pierre (23 Aug 2005) *A new classification framework for urban geospatial Web sites*. Cybergeog: European Journal of Geography [online], (document 318):p 27. URL <http://www.cybergeog.eu/index3115.html>.
- Carpineto**, Claudio & **Romano**, Giovanni (2004) *Concept Data Analysis: Theory and Applications*. John Wiley & Sons, Ltd, ISBN 0-470-85055-8.
- Carr**, Nicholas (6 Aug 2010) *Tracking is an assault on liberty, with real dangers*. The Wall Street Journal. URL <http://www.wsj.com/articles/SB10001424052748703748904575411682714389888>.
- Carroll**, Lewis (1892) *Sylvie and Bruno Concluded*. Macmillan and Co.
- Carter**, Jimmy (15 Jul 2009) *Losing my religion for equality*. The Age. URL <http://www.theage.com.au/federal-politics/losing-my-religion-for-equality-20090714-dk0v.html>.
- Castelein**, Watse, **Grus**, Łukasz, **Crompvoets**, Joep & **Bregt**, Arnold (2010) *A characterization of Volunteered Geographic Information*. In: 13th AGILE International Conference on Geographic Information Science 2010.
- Çöltekin**, Arzu & **Clarke**, Keith C (eds) (Feb 2011a) *Position papers on Virtual Globes or Virtual Geographical Reality: How much detail does a digital earth require? Proceedings of the ASPRS/CaGIS 2010 Workshop, 16 November 2010, Orlando, Florida, USA*.
- Çöltekin**, Arzu & **Clarke**, Keith C (Feb 2011b) *A representation of everything*. In: Çöltekin & Clarke [2011a], pp 9–15.
- Ceglowski**, Maciej (8 Nov 2011) *The social graph is neither*. Pinboard Blog. URL [http://blog.pinboard.in/2011/11/the\\_social\\_graph\\_is\\_neither/](http://blog.pinboard.in/2011/11/the_social_graph_is_neither/).

- Cerf**, Vinton G (29 Sep 2000) *Al Gore and the Internet*. Nettime mailing list. URL <http://amsterdam.nettime.org/Lists-Archives/nettime-1-0009/msg00311>.
- Chan**, K S May (2–8 May 2010a) *Formal methods for Web services: A taxonomic approach*. In: *32nd International Conference on Software Engineering (ICSE'10), Volume 2*, ACM, Cape Town, South Africa, pp 357–360.
- Channel24** (14 May 2015) *A frustrated Toya Delazy leaks her own album online*. Channel24coza. URL <http://www.channel24.co.za/Music/News/A-frustrated-Toya-Delazy-leaks-her-own-album-online-20150514>.
- Chatfield**, Tom (21 Dec 2012) *YouTube: The cult of web video*. BBC — Future. URL <http://www.bbc.com/future/story/20121221-the-cult-of-web-video/>.
- Chaum**, David L (Feb 1981) *Untraceable electronic mail, return addresses, and digital pseudonyms*. *Communications of the ACM*, **24**(2):pp 84–88.
- Chen**, Adrian (26 Nov 2013) *Much ado about Bitcoin*. The New York Times. URL [www.nytimes.com/2013/11/27/opinion/much-ado-about-bitcoin](http://www.nytimes.com/2013/11/27/opinion/much-ado-about-bitcoin).
- Cheru**, Fantu (2012) *African scholars and western Africanists: a world apart*. *Journal of Contemporary African Studies*, **30**(2):pp 193–194.
- Childish**, Billy & **Thomson**, Charles (4 Aug 1999) *The Stuckists: Against conceptualism, hedonism and the cult of the egoist; artist. Manifesto*, The Stuckists. URL <http://stuckism.com/>.
- Chilton**, Steve (5 Sep 2012) *What's Neo?* In: *First International Cartographic Association Neocartography Commission workshop*, London. URL <http://de.slideshare.net/steve8/whats-neo>.
- Christensen**, Nilofer, **Rajabifard**, Abbas & **Paull**, Dan (2014) *Understanding the provision of national location information in Australia: a PSMA case study*. *Journal of Spatial Science*, doi: 10.1080/14498596.2014.880075.
- Cilliers**, Han (Lola) (16 Oct 2014) *Sarkeesian highlights poor handling of terrorist threat by Utah State University*. MWEB GameZone. URL <http://www.mweb.co.za/games/ViewNewsArticle/tabid/2549/Article/16128/Sarkeesian-highlights-poor-handling-of-terrorist-threat-by-Utah-State-University.aspx>.
- Cinnamon**, Jonathan (Mar 2015) *Deconstructing the binaries of spatial data production: Towards hybridity*. *The Canadian Geographer*, **59**(1):pp 35–51, doi: 10.1111/cag.12119.
- Clarke**, Derek G (ed) (2014) *Guidelines of Best Practice for the Acquisition, Storage, Maintenance and Dissemination of Fundamental Geo-Spatial Datasets: Mapping Africa for Africa (MAfA)*. United Nations Economic Commission for Africa (UN ECA). In preparation.
- Clarke**, Derek G, **Cooper**, Antony K, **Liebenberg**, Elri C & **Van Rooyen**, M Hester (Sep 1987) *A national standard for the exchange of digital geo-referenced information*. Special Report SWISK 45, CSIR, Pretoria, South Africa.
- Clarke**, Derek G, **Cooper**, Antony K, **Liebenberg**, Elri C & **Van Rooyen**, M Hester (Mar 1988) *On proposing a national standard for the exchange of digital geo-referenced information*.

## Bibliography

---

- South African Journal of Photogrammetry, Remote Sensing and Cartography, **15**(1):pp 35–41.
- Clarke, Ian, Sandberg, Oskar, Wiley, Brandon & Hong, Theodore W** (2001) *Freenet: A distributed anonymous information storage and retrieval system*. In: *International Workshop on Designing Privacy Enhancing Technologies: Design Issues in Anonymity and Unobservability*, Springer-Verlag New York, Inc, pp 46–66.
- Clear, Martin** (15 Jan 2014) *Why should i reveal my 'real identity' online? anonymity isn't so terrible*. The Guardian. URL <http://www.theguardian.com/commentisfree/2014/jan/15/reveal-real-identity-online-anonymity>.
- Cleophas, Loek, Watson, Bruce W, Kourie, Derrick G, Boake, Andrew & Obiedkov, Sergei** (Dec 2006) *TABASCO: a taxonomy-based domain engineering method*. South African Computer Journal, (37):pp 30–40.
- Clinton, William J** (1994) *Coordinating geographic data acquisition and access: the national spatial data infrastructure*. Executive Order, **12906**:pp 17 671–17 674.
- Cloonan, Michele V & Dove, John G** (4 Jan 2005) *Ranganathan Online: Do digital libraries violate the Third Law?* Library Journal. URL <http://www.libraryjournal.com/article/CA512179.html>.
- Cochrane, Dave** (15 Aug 2014) *Mapping the world all over again*. Wired UK. URL [www.wired.co.uk/news/archive/2014-08/15/mapping-the-world-again](http://www.wired.co.uk/news/archive/2014-08/15/mapping-the-world-again).
- CODATA-ICSTI Task Group on Data Citation Standards and Practices** (13 Sep 2013) *Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data*. Data Science Journal, **12**:pp CIDCR1–CIDCR75.
- Coetzee, Louis & Eksteen, Johan** (2011) *The internet of things — promise for the future? an introduction*. In: **Cunningham, Paul & Cunningham, Miriam** (eds), *IST-Africa 2011 Conference Proceedings*, ISBN 978-1-905824-24-3. URL <http://www.IST-Africa.org/Conference2011/>.
- Coetzee, Louis & Eksteen, Johan** (2012) *Positioning internet of things application, and associated human behavioural changes in a developing context*. In: **Cunningham, Paul & Cunningham, Miriam** (eds), *IST-Africa 2012 Conference Proceedings*, ISBN 978-1-905824-34-2. URL <http://www.IST-Africa.org/Conference2012/>.
- Coetzee, Serena** (ed) (2–4 Oct 2012) *GISSA Ukubuzana 2012 Conference*, Kempton Park, South Africa.
- Coetzee, Serena & Cooper, Antony K** (Nov 2007b) *What is an address in South Africa?* South African Journal of Science, **103**(11/12):pp 449–458.
- Coetzee, Serena & Cooper, Antony K** (10 May 2012) *Opportunities for research and innovation from involvement in standards development — experiences of two researchers*. In: *International Conference on Education in Standardisation (ICES) 2012*, Bali, Indonesia, p 6.
- Coetzee, Serena, Cooper, Antony K, Fleming, Gavin & Netterberg, Inge S J** (2009) *Guest*

- Editorial: Free and Open Source Software for Geospatial (FOSS4G).* South African Computer Journal, **43**:pp 1–2.
- Coetzee, Serena, Cooper, Antony K, Netterberg, Inge & Fleming, Gavin** (eds) (29 Sep – 3 Oct 2008c) *Academic track of the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference, incorporating the GISSA 2008 Conference*, Cape Town, South Africa, ISBN 978-0-620-42177-1. URL <http://conference.osgeo.org/index.php/foss4g/2008/schedConf/presentations>.
- Coetzee, Serena, Cooper, Antony K & Rautenbach, Victoria** (13 Jun 2014) *Part C: Standards for fundamental geo-spatial datasets, in Clarke*. In: Clarke [2014], p 124. ISO/TC 211 document N 3805.
- Coetzee, Serena, Eksteen, Sanet & Grundling, Christopher** (Jun 2013a) *Sustainable development: The contribution from GISc education in South Africa*. South African Journal of Geomatics, **2**(3):pp 246–259.
- Coetzee, Serena, Harvey, Francis, Iwaniak, Adam & Cooper, Antony** (25–30 Aug 2013b) *Sharing and coordinating SDIs in the age of crowdsourcing and mobile technologies*. In: Buchroithner [2013].
- Coetzee, Serena & Smit, Julian** (Aug 2015) *Development of an observatory for spatial planning in south africa: a best practice review*. South African Journal of Geomatics, **4**(3):pp 326–338.
- Coetzee, Serena & Wolff-Piggott, Brendon** (23–28 Aug 2015) *A Review of SDI Literature: Searching for Signs of Inverse Infrastructures*. In: Robbi Sluter *et al* [2015a], pp 113–127. URL <http://www.springer.com/978-3-319-17737-3>.
- Cohen, Jon** (28 Apr 2000) *AIDS researchers decry Mbeki's Views on HIV*. Science, **288**(5466):pp 590–591, doi: 10.1126/science.288.5466.590.
- Coleman, David J, Georgiadou, Yola & Labonte, Jeff** (2009) *Volunteered geographic information: The nature and motivation of producers*. International Journal of Spatial Data Infrastructures Research, Special Issue on GSDI-11, **4**:pp 332–358, doi: 10.2902/1725-0463.2009.04.art16.
- Collins, Katie** (19 Aug 2014b) *You have more to hide in your data trail than you think*. Wired UK. URL <http://www.wired.co.uk/news/archive/2014-08/19/reading-and-interpre...>
- Commonwealth Statisticians (5–9 Sep 2005) *14th Conference of Commonwealth Statisticians*, Cape Town, South Africa. URL <http://www.statssa.gov.za/commonwealth/speakerpresentations.asp>.
- Coomes, Phil** (15 May 2015) *Speakers' corner: The home of free speech*. BBC News. URL <http://www.bbc.com/news/in-pictures-32703071>.
- Cooper, Antony K** (Jul 1986) *National standards for geographical information systems (GIS) data exchange*. In: *Proceedings 1st Computer Science Research Students Conference*, Stutterheim, South Africa, p 15. Also NRIMS CSIR Technical Report TWISK 478.

## Bibliography

---

- Cooper, Antony K** (Sep 1987a) *Geographical information systems*. In: *Proceedings Computer Graphics '87*, Johannesburg, South Africa. Also NRIMS CSIR Technical Report TWISK 542, 10 pp.
- Cooper, Antony K** (Mar 1987b) *Thoughts on exchanging geographical information*. In: *Proceedings 1987 ASPRS-ACSM Annual Convention*, vol 5, Baltimore, MD, USA, pp 1–9. Also NRIMS CSIR Technical Report TWISK 500, 13 pp.
- Cooper, Antony K** (Sep 1988a) *A data structure for exchanging geographical information*. *Quaestiones Informaticae*, 6(2):pp 77–82.
- Cooper, Antony K** (Mar 1988b) *Temporal terrain shading*. In: *Proceedings 1988 ACSM-ASPRS Annual Convention*, vol 2, St Louis, MO, USA, pp 282–290.
- Cooper, Antony K** (Jul 1989a) *Research into the nature of digital geographical information*. In: *Proceedings SAGIS'89*, Pietermaritzburg, South Africa. Also CACDS CSIR Technical Report PKOMP 89/8, 6 pp.
- Cooper, Antony K** (1989b) *A survey of standards for the exchange of digital geo-referenced information*. *South African Journal of Photogrammetry, Remote Sensing and Cartography*, 15(3):pp 136–140.
- Cooper, Antony K** (Jun 1993) *Standards for exchanging digital geo-referenced information*. Master's thesis, University of Pretoria, Pretoria, South Africa.
- Cooper, Antony K** (1997) *South Africa: National Standard for the Exchange of Digital Geo-referenced Information (NES)*. In: Moellering & Hogan [1997], pp 225–235.
- Cooper, Antony K** (Dec 1999) *Halse/Stone/Coleman in the Lady Grey District*. *Familia*, the Quarterly Journal of the Genealogical Society of South Africa, 36(4):pp 158–160.
- Cooper, Antony K** (2 Oct 2003) *Thoughts on categorising bloodstain patterns*. Technical Report 0442-0001-701-A1, CSIR icomtek.
- Cooper, Antony K** (5–9 Sep 2005) *A proposed methodology and infrastructure for standards development and implementation within a national statistical agency*. In: *Commonwealth Statisticians [2005]*, p 9. URL <http://www.statssa.gov.za/commonwealth/speakerpresentations.asp>.
- Cooper, Antony K** (30 Apr 2007) *Participation in the development of standards in ISO/TC 211*. CODI-V pre-conference workshop on the development of an African Metadata Profile, Addis Ababa, Ethiopia.
- Cooper, Antony K** (8 May 2008b) *Some thoughts on humanitarian logistics and quantitative methods*. In: *Humanitarian Logistics: Networks for Africa*, Bellagio, Italy.
- Cooper, Antony K** (27 May 2009a) *African requirements for SDI standardization*. ISO/TC 211 Workshop on Spatial Data Infrastructures (SDIs), Molde, Norway.
- Cooper, Antony K** (28 Apr 2009b) *Geoinformation perspectives on innovation and economic growth*. In: *1st Session of the Committee on Development Information, Science and Technology (CODIST)*, Addis Ababa, Ethiopia. URL <http://researchspace.csir.co.za/dspace/handle/10204/3323>.

- Cooper, Antony K** (1 Jul 2011a) *Geographical information and South Africa's Protection of Information Bill*. In: ICC 2011 Workshop on Open Data Access and Intellectual Property Rights for Cartographers, Richelieu Room, Bibliothèque nationale de France, Paris, France.
- Cooper, Antony K** (5 Sep 2011b) *Unpatriotic clothing*. Business Day. URL <http://www.bdlive.co.za/articles/2011/09/05/unpatriotic-clothing>.
- Cooper, Antony K** (12–13 May 2011c) *Volunteered geographical information and standards for geographical information*. In: OSGeoPL Conference, Wrocław, Poland.
- Cooper, Antony K** (Oct 2013) *Spatial data infrastructures (SDIs) and observatories for spatial analysis*. Spatial Temporal Evidence for Planning South Africa (stepSA), (Policy Note 3). URL [http://stepsa.org/resources/copy\\_of\\_shared-documents/csir-stepsa-policy-note-2-on-sdis-final.pdf/view](http://stepsa.org/resources/copy_of_shared-documents/csir-stepsa-policy-note-2-on-sdis-final.pdf/view).
- Cooper, Antony K** (21–22 Nov 2015) *VGI, crowd-sourcing, citizen science and neogeography are not the same!* In: United Nations Economic Commission for Africa (UN ECA) Expert Group Meeting on Volunteer Geographic Information (VGI), Nairobi, Kenya. A brief presentation for an invited workshop of VGI experts.
- Cooper, Antony K, Byleveld, Piet & Schmitz, Peter M U** (1–4 Dec 2001) *Using GIS to reconcile crime scenes with those indicated by serial criminals*. In: 5th Annual International Crime Mapping Research Conference, Dallas, Texas, USA. URL <http://researchspace.csir.co.za/dspace/handle/10204/2779>.
- Cooper, Antony K & Clarke, Derek G** (1991) *The South African standard for the exchange of digital geo referenced information*. In: Moellering [1991], pp 154–168.
- Cooper, Antony K, Coetzee, Serena, Kaczmarek, Iwona, Kourie, Derrick G, Iwaniak, Adam & Kubik, Tomasz** (31 May– 2 Jun 2011a) *Challenges for quality in volunteered geographical information*. In: Smit [2011].
- Cooper, Antony K, Coetzee, Serena & Kourie, Derrick G** (2010a) *Perceptions of virtual globes, volunteered geographical information and spatial data infrastructures*. *Geomatica*, 64(1):pp 333–348.
- Cooper, Antony K, Coetzee, Serena & Kourie, Derrick G** (2–4 Oct 2012a) *Assessing the quality of repositories of volunteered geographical information*. In: Coetzee [2012].
- Cooper, Antony K, Coetzee, Serena & Kourie, Derrick G** (Apr 2012b) *An assessment of several taxonomies of volunteered geographical information*. In: Díaz et al [2012], pp 21–36.
- Cooper, Antony K, Coetzee, Serena, Rapant, Petr, Laurent, Dominique, Danko, David M, Iwaniak, Adam, Peled, Ammatzia, Moellering, Harold & Düren, Ulrich** (25–30 Aug 2013) *Exploring the impact of a spatial data infrastructure on value-added resellers and vice versa*. In: Buchroithner et al [2013], pp 395–406, doi: 10.1007/978-3-642-32618-9.
- Cooper, Antony K, Coetzee, Serena & Ravno, Kevin** (Feb 2011b) *Some thoughts on geovirtual environments*. In: Çöltekin & Clarke [2011a], pp 17–22.
- Cooper, Antony K & Das, Sonali** (16–22 Aug 2009) *The contribution statistics can make to "strengthening forensic science"*. In: International Statistical Institute [2009].

## Bibliography

- Cooper, Antony K, Dely, Jenny, Gilfillan, T Chris, Kanyama, Leonard & Whisken, Jarrell** (Mar 1996) *Quality assurance: demarcation of the RSA in enumerator areas*. contract report DIS-C 186, CSIR Information Services.
- Cooper, Antony K & Gavin, Elizabeth JO** (2005) *Metadata in Africa and the Middle East*. In: Moellering *et al* [2005], pp 431–450.
- Cooper, Antony K, Gilfillan, T Chris, Kanyama, Leonard, Modise, Lawrence L & Whisken, Jarrell** (Feb 1997) *Quality assurance on the revised demarcation of the RSA in enumerator areas*. contract report DIS-C 214, CSIR Information Services.
- Cooper, Antony K, Hjelmager, Jan N, Nielsen, Anders S & Rapant, Petr** (11–15 Aug 2003a) *A description of spatial data infrastructures (SDIs) using the Unified Modelling Language (UML)*. In: *21st International Cartographic Conference, Durban, South Africa*, ISBN 0-958-46093-0.
- Cooper, Antony K & Hobson, Colin D** (Sep 1991) *GIS education — maximising the effectiveness of available skilled personnel in developing countries*. In: *15th International Cartographic Conference, Bournemouth United Kingdom*.
- Cooper, Antony K, Ittmann, Hans W, Stylianides, Theo & Schmitz, Peter MU** (Dec 2009a) *Ethical issues in tracking cellphones at an event*. OMEGA, Special issue on ethics and operational research, 37(6):pp 1063–1072.
- Cooper, Antony K, Kourie, Derrick G & Coetzee, Serena** (19–21 Oct 2010b) *Thoughts on exploiting instability in lattices for assessing the discrimination adequacy of a taxonomy*. In: *The Seventh International Conference on Concept Lattices and Their Applications (CLA 2010)*, Seville, Spain.
- Cooper, Antony K, Majeke, Bongani, Govender, Sives, Lance, Kate & Tieszen, Larry L** (Nov 2005) *Spatial data content standards for Africa*. In: *AfricaGIS 2005 Conference, Pretoria, South Africa*, p 7. Accessed 8 March 2009 at: URL <http://researchspace.csir.co.za/dspace/handle/10204/1780>.
- Cooper, Antony K, Moellering, Hal, Hjelmager, Jan, Rapant, Petr, Delgado, Tatiana, Laurent, Dominique, Coetzee, Serena, Danko, Dave M, Düren, Ulrich, Iwaniak, Adam, Brodeur, Jean, Abad, Paloma, Huet, Michel & Rajabifard, Abbas** (2012c) *A spatial data infrastructure model from the computational viewpoint*. *International Journal of Geographical Information Science*, 27(6):pp 1133–1151, doi: 10.1080/13658816.2012.741239.
- Cooper, Antony K, Moellering, Harold, Delgado, Tatiana, Düren, Ulrich, Hjelmager, Jan, Huet, Michel, Rapant, Petr, Rajabifard, Abbas, Laurent, Dominique, Iwaniak, Adam, Abad, Paloma & Martynenko, Alexander** (Aug 2007) *An initial model for the computation viewpoint of a spatial data infrastructure*. In: *23rd International Cartographic Conference, Moscow, Russia*.
- Cooper, Antony K, Moellering, Harold, Hjelmager, Jan N, Rapant, Petr, Laurent, Dominique, Abad, Paloma & Danko, David** (Nov 2009b) *Detailed services in a spatial data infrastructure from the computation viewpoint*. In: *24th International Cartographic Conference*.

- ference, Santiago, Chile. URL [http://icaci.org/files/documents/ICC\\_proceedings/ICC2009/html/refer/3\\_9.pdf](http://icaci.org/files/documents/ICC_proceedings/ICC2009/html/refer/3_9.pdf).
- Cooper, Antony K & Nielsen, Anders S** (22 Apr 2000) *Global Spatial Data Infrastructure White Paper: Report of the Exploratory Committee*. White paper, Commission on Spatial Data Standards of the International Cartographic Association. Also MIKOMTEK CSIR report: 0134-0003-704-A1.
- Cooper, Antony K, others (ICA Commission on Incremental Updating & Versioning)** (Aug 2003b) *The concepts of incremental updating and versioning*. In: *21th International Cartographic Conference*, Durban, South Africa, ISBN 0-958-46093-0.
- Cooper, Antony K & Peled, Ammatzia** (Aug 2001) *Incremental updating and versioning*. In: *20th International Cartographic Conference*, vol 4, Beijing, China, pp 2804–2809. URL [http://cartography.tuwien.ac.at/ica/documents/ICC\\_proceedings/ICC2001/icc2001/file/f19007.doc](http://cartography.tuwien.ac.at/ica/documents/ICC_proceedings/ICC2001/icc2001/file/f19007.doc).
- Cooper, Antony K, Rapant, Petr, Hjelmager, Jan, Laurent, Dominique, Iwaniak, Adam, Coetzee, Serena, Moellering, Harold & Düren, Ulrich** (4–8 Jul 2011c) *Extending the formal model of a spatial data infrastructure to include volunteered geographical information*. In: *25th International Cartographic Conference*, Paris, France.
- Cooper, Antony K, Schmitz, Peter M U & Krygsman, Stephan C** (16–19 Aug 2010d) *Tracking cellular telephones to build transport models*. In: *South African Transportation Conference (SATC)*, Pretoria, South Africa.
- Cooper, Antony K, van Huyssteen, Elsona, Das, Sonali, Coetzee, Maria J & Mans, Gerbrand** (May 2014) *Assessment of spatial data infrastructures*. *Town and Regional Planning*, (64):pp 65–75.
- Costa, Hugo, Foody, Giles M, Jiménez, Sílvia & Silva, Luís** (16 Nov 2015) *Impacts of species misidentification on species distribution modeling with presence-only data*. *ISPRS International Journal of Geo-Information*, 4:pp 2496–2518.
- Court of Justice of the European Union** (25 Jun 2013) *Advocate General's Opinion in Case C-131/12, Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González*. PRESS RELEASE No 77/13. InfoCuria — Case-law of the Court of Justice. Advocate General Jääskinen considers that search engine service providers are not responsible, on the basis of the Data Protection Directive, for personal data appearing on Web pages they process.
- Court of Justice of the European Union** (13 May 2014) *Judgement of the Court (Grand Chamber) in Case C-131/12, Google Spain SL, Google Inc v Agencia Española de Protección de Datos (AEPD), Mario Costeja González*. InfoCuria — Case-law of the Court of Justice.
- Cowie, Jim** (19 Nov 2013) *The new threat: Targeted internet traffic misdirection*. Renesys Blog. URL <http://www.renesys.com/2013/11/mitm-internet-hijacking/>.
- Cox, Simon, Schade, Sven & Portele, Clemens** (24 Jun 2010) *Linked data in SDI*. In: *INSPIRE Conference 2010*, Krakow, Poland. URL [http://inspire.jrc.ec.europa.eu/events/conferences/inspire\\_2010/presentations/206\\_pdf\\_presentation.pdf](http://inspire.jrc.ec.europa.eu/events/conferences/inspire_2010/presentations/206_pdf_presentation.pdf).

## Bibliography

---

- Craglia, Max, Goodchild, Michael F, Annoni, Alessandro, Camara, Gilberto, Gould, Michael, Kuhn, Werner, Mark, David, Masser, Ian, Maguire, David, Liang, Steve & Parsons, Ed** (2008) *Next-Generation Digital Earth, a position paper from the Vespucci Initiative for the advancement of Geographic Information Science*. International Journal of Spatial Data Infrastructures Research, 3:pp 146–167.
- Cressey, Daniel** (6 Jul 2012) *Nature Publishing Group wins long-running libel trial*. Nature, doi: 10.1038/nature.2012.10965. URL <http://www.nature.com/news/nature-publishing-group-wins-long-running-libel-trial-1.10965>.
- Cross, Michael** (7 Jun 2007) *Address database plan finally abandoned*. The Guardian. URL <http://www.guardian.co.uk/technology/2007/jun/07/guardianweeklytechnologysection.freeourdata>.
- Crown, The** (1872 (35 and 36 Vict)) *Parks Regulation Act 1872*. C 15, UK. This is the amended version of the Act — really need to see the original and get a better URL, URL <http://www.legislation.gov.uk/ukpga/Vict/35-36/15/contents>.
- Crown, The** (1997) *The Royal Parks and Other Open Spaces Regulations*. SI 1997/1639, UK. URL <http://www.opsi.gov.uk/si/si1997/19971639.htm>.
- CSI** (27 Mar 2014) *The South African Geo-Information Management Strategy (SAGIMS): A draft framework to guide the development of the strategy*. Tech rep, Committee for Spatial Information. For discussion purposes, URL <http://www.sasdi.gov.za/DCPRDocs/Draft Strategy Framework 27 March 2014 version 1.pdf>.
- CSIR** (Oct 2008) *Earth observation for Africa — eyes in the sky*. Pamphlet, CSIR, Pretoria, South Africa.
- Dahlgren, Peter** (20 Sep 2012) *Public intellectuals, online media, and public spheres: Current realignments*. International Journal of Politics, Culture, and Society, 25:pp 95–110.
- Dale, Peter F** (1991) *Land information systems*. In: **DJ, Maguire, MF, Goodchild & DW, Rhind** (eds), *Geographical information systems, Volume 2: Applications*, Longman Scientific & Technical, London, pp 85–99.
- Daly, Patrick W** (8 Sep 2003) *Customizing Bibliographic Style Files: This paper describes program makebst version 4.1 from 2003/09/08*. Including additions by Arthur Ogawa, [ogawa@teleport.com](mailto:ogawa@teleport.com).
- Daly, Patrick W** (6 Sep 2006) *Natural Sciences Citations and References (Author-Year and Numerical Schemes): This paper describes package natbib version 7.4a from 2006/09/06*.
- Davies, Dai** (27–30 Oct 2002) *GÉANT — The next generation of European backbone networking*. In: *Fall 2002 Internet2 Member Meeting*, Los Angeles, CA, USA.
- Davis, Marc, King, Simon, Good, Nathan & Sarvas, Risto** (10–16 Oct 2004) *From context to content: Leveraging context to infer media metadata*. In: *12th Annual ACM International Conference on Multimedia (MM'04)*, ACM, New York, NY, USA, pp 188–195.
- Dawkins, Richard** (9 Jul 1998) *Postmodernism disrobed*. Nature, 394:pp 141–143. URL <http://old.richarddawkins.net/articles/824-postmodernism-disrobed>.

- Dawood, Ayesha** (1 Oct 2014a) *Blog: Broadband policy reflects state's failure*. Business Day Live. URL <http://www.bdlive.co.za/blogs/.../blog-broadband-policy-reflects-states-failure>.
- Dawsey, Joshua** (18 Dec 2012) *Youtube's top ten viral videos of 2012*. The Wall Street Journal. URL <http://blogs.wsj.com/speakeasy/2012/12/18/youtubes-top-ten-viral-videos-of-2012/>.
- DDI (Oct 2009) *Data Documentation Initiative (DDI) Technical Specification*. DDI Alliance. URL <http://www.ddialliance.org/>.
- De Long, J Bradford & Lang, Kevin** (Dec 1992) *Are all economic hypotheses false?* Journal of Political Economy, **100**(6):pp 1257–1272.
- De Longueville, Bertrand, Ostländer, Nicole & Keskitalo, Carina** (2010b) *Addressing vagueness in Volunteered Geographic Information (VGI) — A case study*. International Journal of Spatial Data Infrastructures Research, Special Issue on GSDI-11, **5**.
- de Montjoye, Yves-Alexandre, Hidalgo, César A, Verleysen, Michel & Blondel, Vincent D** (25 Mar 2013) *Unique in the crowd: The privacy bounds of human mobility*. Scientific Reports, **3**(1376):pp 1–5, doi: 10.1038/srep01376.
- Debord, Guy-Ernest** (1955) *Introduction to a critique of urban geography*. Les Lèvres Nues, (6).
- Dessers, Ezra** (2012) *Spatial Data Infrastructures at work: A comparative case study on the spatial enablement of public sector processes*. PhD thesis, Katholieke Universiteit Leuven, Leuven, België.
- Devillers, Rodolphe, Bégin, Daniel & Vandecasteele, Arnaud** (18 Sep 2012) *Is the rise of volunteered geographic information (VGI) a sign of the end of national mapping agencies as we know them?* In: GIScience 2012 workshop “Role of Volunteer Geographic Information in Advancing Science: Quality and Credibility”, Columbus, OH, USA.
- Dewey, Caitlin** (30Dec 2013) *A year of Internet hoaxes: Fake events but very real coverage*. The Washington Post. URL <https://www.washingtonpost.com/news/arts-and-entertainment/wp/2013/12/30/a-year-of-internet-hoaxes-fake-events-but-very-real-coverage/>.
- Dewey, Caitlin** (18May 2015a) *If you could print out the whole Internet, how many pages would it be?* The Washington Post. URL <https://www.washingtonpost.com/news/the-intersect/wp/2015/05/18/if-you-could-print-out-the-whole-internet-how-many-pages-would-it-be/>.
- Dewey, Caitlin** (22Dec 2015b) *What happened to the 15 people the internet hated most in 2015*. The Washington Post. URL <https://www.washingtonpost.com/news/the-intersect/wp/2015/12/22/whatever-happened-to-the-15-people-the-internet-hated-most-in-2015/>.
- Díaz, Laura, Granell, Carlos & Huerta, Joaquin** (eds) (Apr 2012) *Discovery of Geospatial Resources: Methodologies, Technologies, and Emergent Applications*. IGI Global, ISBN 978-1-4666-0945-7, 355 pp.

## Bibliography

---

- Dillman**, Don A (Mar–Apr 2008) *Comment: Errors Galore*. *Interfaces*, 38(2):pp 133–134.
- Ding**, Li, **Peristeras**, Vassilios & **Hausenblas**, Michael (May/Jun 2012) *Linked open government data: Guest editors' introduction*. *IEEE Intelligent Systems*:pp 11–15.
- Diu**, Nisha Lilia (16 Apr 2015) *'enough with the catcalls': Meet the woman making a street harassment map*. *The Telegraph*. URL <http://www.telegraph.co.uk/women/womens-life/11520753/Everyday-Sexism-Meet-the-Brazilian-woman-mapping-street-harrassment.html>.
- Dlodlo**, Nomusa, **Mbecke**, Paul, **Mofolo**, Mofolo & **Mhlanga**, Martin (12–14 Dec 2013) *The internet of things in community safety and crime prevention for South Africa*. In: *International Joint Conferences on Computer, Information and Systems Sciences and Engineering (CISSE 2013)*, University of Bridgeport, CT, USA, online.
- Dobson**, Jerome E & **Fisher**, Peter F (Spring 2003) *Geoslavery*. *IEEE Technology and Society Magazine*:pp 47–52.
- Doctorow**, Cory (26 Aug 2001) *Metacrap: Putting the torch to seven straw-men of the meta-utopia*. online posting. Accessed 4 July 2014, URL <http://www.well.com/~doctorow/metacrap.htm>.
- Doctorow**, Cory (2 Sep 2009) *Not every cloud has a silver lining*. *The Guardian*. URL <http://www.guardian.co.uk/technology/2009/sep/02/cory-doctorow-cloud-computing>.
- Dorn**, Helen, **Törnros**, Tobias & **Zipf**, Alexander (Sep 2015) *Quality evaluation of VGI using authoritative data — a comparison with land use data in Southern Germany*. *ISPRS International Journal of Geo-Information*, 4:pp 1657–1671.
- Downes**, Larry (23 Dec 2015) *The 4 worst patents of 2015*. *The Independent*. URL <https://www.washingtonpost.com/news/innovations/wp/2015/12/14/the-4-worst-patents-of-2015/>.
- Du Plooy**, Niell (Nov 2012) *Assessment of OpenStreetMap data quality across different urban hierarchic levels: A comparative study*. Bsc honours report, Department of Computer Science, University of Pretoria. Unpublished.
- Dun**, Twll (8 Jan 2016) *Stop people on the Internet setting up frankly barking online petitions*. *Change.org*. URL <https://www.change.org/p/david-cameron-mp-jeremy-corbyn-mp-stop-people-on-the-internet-setting-up-frankly-barking-online-petitions>.
- Dunbar**, Robin (2010) *How many friends does one person need? Dunbar's number and other evolutionary quirks*. Faber and Faber Limited, London, ISBN 978-0-571-25342-5.
- Duncan**, Jane (13 Jun 2011) *The Prevention of Scholarship Bill*. *The Media Online*. URL <http://themediainline.co.za/2011/06/the-prevention-of-scholarship-bill/>.
- Dunsire**, Gordon (12–14 Feb 2013) *Granularity in library linked open data*. In: *Code4Lib 2013*, Chicago, USA. Keynote presentation.

- Duval, E, Hodgins, W, Sutton, S & Weibel, S** (2002) *Metadata Principles and Practicalities*. D-Lib magazine, 8(4).
- Economist**, The (28 Sep 2013) *Looks good on paper: A flawed system for judging research is leading to academic fraud*. The Economist: Scientific research. URL <http://www.economist.com/node/21586845/>.
- Edsall, Robert M, Barbour, Laura & Hoffman, Johanna** (23–28 Aug 2015) *Complementary Methods for Citizen Mapping of Ecosystem Services: Comparing Digital and Analog Representations*. In: Robbi Sluter et al [2015a], pp 295–307. URL <http://www.springer.com/978-3-319-17737-3>.
- EE Publishers (Jul 2013) *SANSA to acquire Spot 6 imagery of South Africa*. EngineerIT. URL <http://eepublishers.co.za/printarticle/view?sid=36483>.
- Eklund, Pieta** (24 Oct 2012) *Open access and predatory publishers: A guide to reviewing open access journals*. Tech rep, University of Borås, Borås, Sweden. URL <http://bada.hb.se/handle/2320/11421/>.
- Electronic Frontier Foundation** (2012) *Anti-Counterfeiting Trade Agreement: What is ACTA?* Electronic Frontier Foundation. URL <https://www.eff.org/issues/acta>.
- Ellul, Claire, Winer, Daniel, Mooney, John & Foord, Jo** (2012) *Bridging the Gap between Traditional Metadata and the Requirements of an Academic SDI for Interdisciplinary Research*, chap 4. In: Rajabifard & Coleman [2012], pp 57–77.
- Elmer-Dewitt, Phillip** (6 Dec 1993) *First nation in cyberspace*. TIME International, (49). URL <http://www.chemie.fu-berlin.de/outerspace/internet-article.html>.
- Elphinstone, Chris D, Potgieter, T Chris, Melinda A amd Gilfillan, Leslie, C, Schmitz, Peter MU, Cooper, Antony K & Pols, A** (Jul 1999) *Quality assurance of the police stations and their areas of jurisdiction*. In: *Earth Data Information Systems Conference (EDIS '99)*, Pretoria, South Africa.
- Elwood, Sarah** (2008a) *Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS*. GeoJournal, 72:pp 173–183.
- Elwood, Sarah** (2008b) *Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice*. GeoJournal, 72:pp 133–135.
- Elwood, Sarah, Goodchild, Michael F & Sui, Daniel Z** (2012) *Researching volunteered geographic information: Spatial data, geographic research, and new social practice*. Annals of the Association of American Geographers, doi: 10.1080/00045608.2011.595657. In press.
- Engelhardt, Tom** (12 Nov 2013) *The nsa mistakes omniscience for omnipotence*. The Nation. URL <http://www.thenation.com/article/177123/nsa-mistakes-omniscience-omnipotence>.
- Epstein, Robert & Robertson, Ronald E** (4 Aug 2015) *The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections*. PNAS:pp E4512–E4521. URL [www.pnas.org/cgi/doi/10.1073/pnas.1419828112](http://www.pnas.org/cgi/doi/10.1073/pnas.1419828112).

## Bibliography

---

- Erickson, J D** (1984) *The LACIE experiment in satellite aided monitoring of global crop production*. In: **Woodwell, G M** (ed), *The Role of Terrestrial Vegetation in the Global Carbon Cycle: Measurement by Remote Sensing*, John Wiley & Sons, Ltd.
- European Parliament** (2003) *Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information*. OJ L 345/90, European Union. Accessed 9 March 2009, URL <http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:EN:PDF>.
- European Parliament** (2007) *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*. OJ L 108/1, European Union. Accessed 9 March 2009, URL <http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:EN:PDF>.
- Evans, Jenni** (7 May 2015) *Media moguls take copyright battle to court*. Fin24. URL <http://www.fin24.com/Companies/ICT/Media-moguls-take-copyright-battle-to-court-20150506>.
- Evans, Jenni** (14 Jan 2016) *Drought-stricken areas runneth over thanks to massive water drive*. News24. URL <http://www.news24.com/SouthAfrica/News/drought-stricken-areas-runneth-over-thanks-to-massive-water-drive-20160114>.
- “expedition”** (4 Oct 2009) *Somali pirate coast*. Blog with KML. Google Earth Community Forum. This article has now been incorporated into the Google Earth Community “Somalia Pirate Attacks”. Accessed 4 July 2014, URL <https://productforum.google.com/forum/#!msg/gec-travel-information-moderated/3tXTwjwhujE/HsykpBT71kQJ/>.
- Eyres, Harry** (24 Aug 2012) *Maps of the human art*. Financial Times. URL <http://www.ft.com/eyres>.
- Fahmy, Amel, Abdelmonem, Enas, Angieand Hamdy & Badr, Ahmed** (2014) *Towards a safer city. sexual harassment in Greater Cairo: Effectiveness of crowdsourced data*. Research report Dep. No: 2014/13131, HarassMap, in collaboration with the Youth and Development Consultancy Institute (Etijah) and the International Development Research Center (IDRC). URL [http://harassmap.org/en/wp-content/uploads/2013/03/Towards-A-Safer-City\\_full-report\\_EN-.pdf](http://harassmap.org/en/wp-content/uploads/2013/03/Towards-A-Safer-City_full-report_EN-.pdf).
- Fairweather, Alistair** (9 Sep 2013) *How the NSA sabotaged the internet*. Mail & Guardian Online. URL <http://mg.co.za/2013-09-09-how-the-nsa-has-sabotaged-the-entire-internet>.
- Fanelli, Daniele** (2009) *How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data*. PLoS ONE, **4**(5):p e5738, doi: 10.1371/journal.pone.0005738.
- Fang, Ferric C, Steen, R Grant & Casadevall, Arturo** (2012) *Misconduct accounts for the majority of retracted scientific publications*. Proceedings of the National Academy of Sciences of the United States of America, **109**(42):pp 17 028–17 033.
- Faris, Robert & Villeneuve, Nart** (2008) *Measuring global Internet filtering*. Cambridge:

- MIT Press, pp 5–28. Deibert, Ronald and Palfrey, John and Rohozinski, Rafal and Zittrain, Jonathan and Stein, Janice Gross (eds).
- Farivar**, Cyrus (25 Aug 2014) *NSA built “Google-like” interface to scan 850+ billion meta-data records*. Ars Technica. URL <http://arstechnica.com/tech-policy/2014/08/nsa-built-google-like-interface-to-scan-850-billion-metadata-records/>.
- Farwell**, Lawrence, **Richardson**, Drew C & **Richardson**, Graham M (2013) *Brain fingerprinting field studies comparing p300-mermer and p300 brainwave responses in the detection of concealed information*. Cognitive Neurodynamics, 7(4):pp 263–299.
- Fawkes**, Guido (27 Apr 2011) *Silencing court jesters*. Guy Fawkes’ blog. URL <http://order-order.com/2011/04/27/silencing-court-jesters/>.
- Feffer**, John (12 Mar 2010) *Will Facebook remake the world?* Harvard International Review. URL <http://hir.harvard.edu/will-facebook-remake-the-world>.
- Fenton-Smith**, Richard (13 Mar 2015) *Can soup change the world?* BBC News: Magazine. URL <http://www.bbc.com/news/magazine-31594513>.
- FGDC (1998) *Content Standard for Digital Geospatial Metadata*. FGDC-STD-001-1998, Federal Geographic Data Committee (FGDC), Washington, DC, United States of America.
- FGDC (2006) *FGDC – ISO Crosswalk v4.1*. Crosswalk, Federal Geographic Data Committee (FGDC), Washington, DC, United States of America. Developed by Intergraph Corporation, under contract, URL [http://www.fgdc.gov/metadata/documents/FGDC\\_Sections\\_v40.xls](http://www.fgdc.gov/metadata/documents/FGDC_Sections_v40.xls).
- FIG (11–16 Apr 2010) *FIG Congress 2010: Facing the Challenges — Building the Capacity*, Sydney, Australia.
- Fildes**, Jonathan (19 Aug 2009) *Web tool oversees Afghan election*. BBC News. URL <http://news.bbc.co.uk/go/pr/fr/-/2/hi/technology/8209172.stm>.
- Finkelstein**, Sydney (26 Mar 2015) *Will there be a revolt in the sharing economy?* BBC — Capital. URL <http://www.bbc.com/capital/story/20150325-revolt-in-the-sharing-economy>.
- Fischer-Baum**, Reuben & **Bialik**, Carl (13 Oct 2015) *Uber is taking millions of Manhattan rides away from taxis*. FiveThirtyEight. URL <http://fivethirtyeight.com/features/uber-is-taking-millions-of-manhattan-rides-away-from-taxis/>.
- Fiveash**, Kelly (12 Mar 2014) *Five unbelievable headlines that claim Tim Berners-Lee ‘invented the internet’*. The Register. URL [http://www.theregister.co.uk/2014/03/12/tim\\_berniers\\_lee\\_inventor\\_of\\_the\\_internet\\_headline\\_fails/](http://www.theregister.co.uk/2014/03/12/tim_berniers_lee_inventor_of_the_internet_headline_fails/).
- FNC (24 Oct 1995) *FNC Resolution: Definition of “Internet”*. Federal Networking Council. URL [http://www.itrd.gov/fnc/Internet\\_res.html](http://www.itrd.gov/fnc/Internet_res.html).
- Foer**, Jonathan Safran (8 Jun 2013) *How not to be alone*. The New York Times Sunday Review: The opinion pages. URL <http://www.nytimes.com/2013/06/09/opinion/sunday/how-not-to-be-alone.html?src=me&ref=general>.

## Bibliography

---

- Foong, Cheryl** (2010) *Open content licensing of public sector information and the risk of tortious liability for australian governments*. eLaw Journal, **17**(2):pp 23–49. URL <http://eprints.qut.edu.au/42572/>.
- Foster, Alec & Dunham, Ian M** (2014) *Volunteered geographic information, urban forests, & environmental justice*. Computers, Environment and Urban Systems, doi: 10.1016/j.compenvurbsys.2014.08.001.
- Foucault, Michel** (Summer 1982) *The subject and power*. Critical Inquiry, **8**(4):pp 777–795.
- Foucault, Michel** (Autumn 2008) “panopticism” from “discipline & punish: The birth of the prison”. Race/Ethnicity: Multidisciplinary Global Contexts, **2**(1):pp 1–12. English Translation by Alan Sheridan, 1977 (New York: Pantheon). Originally published in French in 1975 as *Surveiller et Punir* (Paris: Editions Gallimard), URL <http://www.jstor.org/stable/25594995>.
- Fowler, H W & Gowers, Ernest** (1965) *A dictionary of modern English usage*. Oxford University Press, 2nd edn. Reprinted with corrections, 1983.
- Freeman, Hadley** (26 Dec 2009) *The baiting and the ‘snark’*. Weekend Post:p 8.
- Frege, Gottlob** (1892) *Über Sinn und Bedeutung*. Zeitschrift für Philosophie und philosophische Kritik, **100**:pp 25–50. Siehe auch Wikipedia: Über Sinn und Bedeutung, URL [https://www.tu-chemnitz.de/phil/english/sections/linguist/independent/kursmaterialien/logling/frege\\_sinnundbedeut.pdf](https://www.tu-chemnitz.de/phil/english/sections/linguist/independent/kursmaterialien/logling/frege_sinnundbedeut.pdf), <http://www.uni-konstanz.de/philosophie/files/frege.pdf>.
- Friedersdorf, Conor** (Jul 2014) *The Latest Snowden Leak Is Devastating to NSA Defenders*. The Atlantic. URL <http://www.theatlantic.com/politics/archive/2014/07/the-latest-snowden-leak-is-devastating-to-nsadefenders/373991/>.
- Friedland, Julian** (2005) *The utility of offshoring: A Rawlsian Critique*. Electronic Journal of Business Ethics and Organization Studies, **10**(1).
- Frost, Amber** (27 Jul 2015) *The pornographic propaganda that was used against marie antoinette*. Dangerous Minds. URL [http://dangerousminds.net/comments/the\\_pornographic\\_propaganda\\_that\\_was\\_used\\_against\\_marie\\_antoinette](http://dangerousminds.net/comments/the_pornographic_propaganda_that_was_used_against_marie_antoinette).
- Galton, Francis** (1907) *Vox populi (the wisdom of crowds)*. Nature, **75**(1949):pp 450–451. Republished online with follow-up letters by The Wisdom of Crowds, URL <http://wisdomofcrowds.blogspot.co.za/2009/12/vox-populi-sir-francis-galton>.
- Ganning, Joanna P, Coffin, Sarah L, McCall, Benjamin & Carson, Kathleen** (2014) *Goals, challenges, and capacity of regional data portals in the United States: An updated understanding of long-standing discussions*. Journal of Urban Technology, **21**(4):pp 125–139.
- Ganter, Bernhard** (1 Oct 2007) *fca.sty L<sup>A</sup>T<sub>E</sub>X — macros for Formal Concept Analysis*. TU Dresden.
- Ganter, Bernhard & Wille, Rudolf** (1997) *Applied lattice theory: Formal concept analysis*. Preprints, URL <http://wwwbib.mathematik.tudarmstadt.de/Math-Net/Preprints/Listen/pp97.html>.

- Garber, Megan** (Sep 2014) *How to stop a rumor online (before the rumor becomes a lie)*. The Atlantic. URL <http://www.theatlantic.com/technology/archive/2014/09/how-to-stop-a-rumor-online-before-the-rumor-becomes-a-lie/380778/>.
- Garratt, Rod & Hayes, Rosa** (28 Aug 2015) *Entry and exit leads to zero profit for Bitcoin miners*. Liberty Street Economics. URL <http://libertystreeteconomics.newyorkfed.org/2015/08/entry-and-exit-leads-to-zero-profit-for-bitcoin-miners.html#.Vq0VPC4lnIU>.
- Gazan, Rich** (Nov/Dec 2008) *Social annotations in digital library collections*. D-Lib Magazine, **14**(11/12).
- Gellman, Barton, Tate, Julie & Soltani, Ashkan** (5 Jul 2014) *In NSA-intercepted data, those not targeted far outnumber the foreigners who are*. The Washington Post. URL [http://www.washingtonpost.com/world/national-security/in-nsa-intercepted-data-those-not-targeted-far-outnumber-the-foreigners-who-are/2014/07/05/8139adf8-045a-11e4-8572-4b1b969b6322\\_story.html](http://www.washingtonpost.com/world/national-security/in-nsa-intercepted-data-those-not-targeted-far-outnumber-the-foreigners-who-are/2014/07/05/8139adf8-045a-11e4-8572-4b1b969b6322_story.html).
- George, Jongikaya Aubrey** (Oct 2010) *Efforts and activities to make data accessible through development of national geospatial data clearinghouse*. In: GSDI-12, Singapore.
- Georgiadou, Yola, Bana, Benson, Becht, Robert, Hoppe, Robert, Ikingura, Justinian, Kraak, Menno-Jan, Lance, Kate, Lemmens, Rob, Hemed Lungo, Juma, McCall, Michael, Miscione, Gianluca & Verplanke, Jeroen** (2011) *Sensors, empowerment, and accountability: a Digital Earth view from East Africa*. International Journal of Digital Earth, **4**(4):pp 285–304.
- Georgiadou, Yola, Lungo, J H & Richter, C** (2013) *Citizen sensors or extreme publics? Transparency and accountability interventions on the mobile geoweb*. International Journal of Digital Earth, doi: 10.1080/17538947.2013.782073. In press.
- Gervais, Daniel J** (2009) *The Tangled Web of UGC: Making Copyright Sense of User-Generated Content*. Vanderbilt Journal of Entertainment and Technology Law, **11**(4):pp 841–870. URL [http://works.bepress.com/daniel\\_gervais/17](http://works.bepress.com/daniel_gervais/17).
- Giles, Jim** (Dec 2005) *Internet encyclopedias go head to head*. Nature, **438**(7070):pp 900–901. URL <http://www.nature.com/nature/journal/v438/n7070/full/438900a.html>.
- GIM International** (14 Sep 2010) *C. Dana Tomlin to Enter URISA's GIS Hall of Fame*. GIM International. URL <http://www.gim-international.com/content/news/c-dana-tomlin-to-enter-urisa-s-gis-hall-of-fame-2>.
- Ginsburg, Isaac** (10 Dec 2001) *The disregard syndrome: A menace to honest science?* The Scientist. URL <http://www.the-scientist.com/?articles.view/articleNo/13745/>.
- Girres, Jean-François & Touya, Guillaume** (2010) *Quality assessment of the French Open-StreetMap dataset*. Transactions in GIS, **14**(4):pp 435–459.
- Giuliani, Gregory, Guigoz, Yaniss, Lacroix, Pierre, Ray, Nicolas & Lehmann, Anthony** (2016) *Facilitating the production of ISO-compliant metadata of geospatial datasets*. International Journal of Applied Earth Observation and Geoinformation, **44**:pp 239–243.

## Bibliography

---

- Gladstone**, Narick (31 Mar 2015) *E-vigilantes take on Islamic State*. Business Day. URL <http://www.bdlive.co.za/world/2015/03/31/e-vigilantes-take-on-islamic-state>.
- Glanz**, James, **Rotella**, Sebastian & **Sanger**, David E (21 Dec 2014) *In 2008 Mumbai attacks, piles of spy data, but an uncompleted puzzle*. New York Times. URL <http://www.nytimes.com/2014/12/22/world/asia/in-2008-mumbai-attacks-piles-of-spy-data-but-an-uncompleted-puzzle>.
- GMWatch** (30 Nov 2013) *Conflicts of interest at Food and Chemical Toxicology and Elsevier*. GMWatch. URL <http://www.gmwatch.org/news/archive/2013/15193-conflicts-of-interest-at-food-and-chemical-toxicology-and-elsevier>.
- Goldberg**, Adrian (3 Feb 2012) *The dark web: Guns and drugs for sale on the internet's secret black market*. BBC News: Business. URL <http://www.bbc.co.uk/news/business-16801382>.
- Golder**, Scott A & **Huberman**, Bernardo A (2006) *Usage patterns of collaborative tagging systems*. Journal of Information Science, **32**(2):pp 198–208.
- Goldman**, Alex (20 Sep 2013) *The breaking news consumer's handbook*. On The Media. URL [http://www.onthemedial.org/tags/breaking\\_news\\_consumers\\_handbook/](http://www.onthemedial.org/tags/breaking_news_consumers_handbook/).
- Goodchild**, Michael F (2006) *Commentary: GIScience ten years after ground truth*. Transactions in GIS, **10**(5):pp 687–692.
- Goodchild**, Michael F (2007b) *Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0*. International Journal of Spatial Data Infrastructures Research, **2**:pp 24–32. Editorial.
- Goodchild**, Michael F (2008b) *Spatial accuracy 2.0*. In: Zhang & Goodchild [2008], pp 1–7.
- Goodchild**, Michael F (Mar 2008c) *The use cases of digital earth*. International Journal of Digital Earth, **1**(1):pp 31–42.
- Goodchild**, Michael F & **Hill**, L (2008) *Introduction to digital gazetteer research*. International Journal of Geographical Information Science, **22**(10):pp 1039–1044.
- Google Search Engine Optimization Starter Guide (2010) *Google Search Engine Optimization Starter Guide*. Google Inc. Accessed 2015-05-24.
- Google terms of service (2010) *Google Maps/Earth Terms of Service*. Google. Accessed 30 Jan 2010, URL [http://maps.google.com/help/terms\\_maps.html](http://maps.google.com/help/terms_maps.html).
- Gopnik**, Adam (19 Apr 2013) *Dzhokhar Tsarnaev, Lost and Found*. The New Yorker. URL <http://www.newyorker.com/online/blogs/newsdesk/2013/04/dzhokhar-tsarnaev-lost-and-found>.
- Gordon**, Anna (13 Aug 2014a) *Humanizing big data: The everyday impact of crowd science*. VentureBeat. URL <http://venturebeat.com/2014/08/13/humanizing-big-data-the-everyday-impact-of-crowd-science/>.

- Gordon, Anna** (14 May 2014b) *Inside the mind of a crowd scientist, an emerging flavor of data scientist*. VentureBeat. URL <http://venturebeat.com/2014/05/14/crowd-scientist/>.
- Gorton, Kristyn & Garde-Hansen, Joanne** (12 Apr 2012) *From old media whore to new media troll*. Feminist Media Studies, doi: 10.1080/14680777.2012.678370. Preview View, URL <http://www.tandfonline.com/doi/full/10.1080/14680777.2012.678370>.
- Govender, Nicole** (Nov 2011) *An analysis of quality of volunteered geographic information in OpenStreetMap*. Bsc honours report, Department of Computer Science, University of Pretoria. Unpublished.
- Graham, Mark** (2 Dec 2009) *Wikipedia's known unknowns*. Guardian Online. URL <http://www.guardian.co.uk/technology/2009/dec/02/>.
- Graham, Mark** (Sep 2010) *Neogeography and the Palimpsests of Place: Web 2.0 and the Construction of a Virtual Earth*. Tijdschrift voor Economische en Sociale Geografie, **101**(4):pp 422–436. URL <http://www.wiley.com/bw/journal.asp?ref=0040-747x>.
- Granick, Jennifer** (24 Jun 2013) *Surveillance myth #1: I have nothing to hide*. Center for Internet and Society Blog. URL <http://cyberlaw.stanford.edu/blog/2013/06/surveillance-myth-1-i-have-nothing-hide>.
- Grant, Bob** (7 May 2009a) *Elsevier published 6 fake journals*. The Scientist. URL <http://www.the-scientist.com/?articles.view/articleNo/27383/title/Elsevier-published-6-fake-journals/>.
- Grant, Bob** (30 Apr 2009b) *Merck published fake journal*. The Scientist. URL <http://www.the-scientist.com/blog/display/55671/>.
- Gray, Eve** (13 Jul 2009a) *Intellectual property: Clash gets to heart of how to build a 'better life for all'*. Business Day:p 7.
- Gray, Richard** (2 Jan 2015) *Crowdfunded science: harnessing the wisdom of the crowd, or selling out?* The Guardian. URL <http://www.theguardian.com/science/2015/jan/02/crowdfunded-science-scientists-fund-research>.
- Greenwald, Glenn** (6 Jun 2013) *NSA collecting phone records of millions of Verizon customers daily*. The Guardian. URL <http://www.guardian.co.uk/world/2013/jun/06/nsa-phone-records-verizon-court-order/>.
- Grieneisen, Michael L & Zhang, Minghua** (Oct 2012) *A comprehensive survey of retracted articles from the scholarly literature*. PLoS ONE, **7**(10):p e44118, doi: 10.1371/journal.pone.0044118.
- Griffin, Andrew** (8 Jan 2016) *UK spying laws criticised by Government's own watchdog as Theresa May's claim that Snoopers' Charter doesn't block encryption is scrutinised*. The Independent. URL <http://www.independent.co.uk/life-style/gadgets-and-tech/news/uk-spying-laws-criticised-by-government-s-own-watchdog-as-theresa-may-s-claim-that-snoopers-charter-a6802986>.

## Bibliography

---

- gROADS (2014) *gROADS: Global Roads Data. Home page*. Global Roads Data Development Task Group of CODATA, the Committee on Data for Science and Technology of the International Council for WScience (ICSU). URL <https://ciesin.columbia.edu/confluence/display/roads/Global+Roads+Data>.
- Gross, Doug (22 Oct 2014) *40% of Web users have been harassed, says survey*. CNN. URL <http://edition.cnn.com/2014/10/22/tech/web/trolling-online-abuse/index.html?eref=edition>.
- Guélat, Jean-Christophe (20–21 Aug 2009) *Integration of user generated content into national databases — revision workflow at swisstopo*. In: *1st EuroSDR Workshop on Crowd Sourcing for Updating National Databases*, Wabern, Switzerland. URL [http://www.eurosd.net/workshops/crowdsourcing\\_2009/](http://www.eurosd.net/workshops/crowdsourcing_2009/).
- Guillén, Mauro F & Suárez, Sandra L (2005) *Explaining the global digital divide: Economic, political and sociological drivers of cross-national internet use*. *Social forces*, **84**(2):pp 681–708.
- Guptill, Steve C & Morrison, Joel L (1995) *Elements of spatial data quality*. Elsevier and the International Cartographic Association.
- Gush, Kim, Cambridge, Grant & Smith, Ronel (2004) *The Digital Doorway — minimally invasive education in Africa*. In: *ICT in Education Conference*.
- Gutierrez, Peter (16 Apr 2014) *At ENC 2014: A GNSS Wake Up Call for Europe. Speakers criticize European leaders' failure to address issue of GNSS vulnerability*. Inside GNSS News. URL <http://www.insidegnss.com/node/3985>.
- Gutmann, Myron P & Stern, Paul C (eds) (2007) *Putting people on the map: Protecting confidentiality with linked social-spatial data*. National Academies Press, ISBN 978-0-309-10414-2. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data, Committee on the Human Dimensions of Global Change, Division of Behavioral and Social Sciences and Education.
- Gyamfi-Aidoo, Jacob, Schwabe, Craig & Govender, Sives (Feb 2006) *Determination of the Fundamental Geo-spatial Datasets for Africa through a User Needs Analysis*. Human Sciences Research Council, Pretoria, South Africa, 373 pp.
- Haden, David (9 Jul 2008) *A short enquiry into the origins and uses of the term "neogeography"*. D'log. Accessed 4 July 2014, URL <http://www.d-log.info/on-neogeography.pdf>.
- Haines, Lester (16 Jul 2006) *Chinese black helicopters circle Google Earth: Mystery military project wows the crowd*. The Register. URL [http://www.theregister.co.uk/2006/07/19/huangyangtan\\_mystery/print.html](http://www.theregister.co.uk/2006/07/19/huangyangtan_mystery/print.html).
- Haklay, Mordechai (2010) *How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets*. *Environment and Planning B: Planning and Design*, **37**:pp 682–703.
- Hambridge, Sally (Oct 1995) *Netiquette guidelines*. Request For Comments 1855, Internet Engineering Task Force (IETF). URL <http://tools.ietf.org/html/rfc1855>.

- Hand, John** (23 Oct 2012) *Ceefax service to end after 38 years on BBC*. BBC News. URL <http://www.bbc.co.uk/news/uk-20032882>.
- Haney, CD** (22 May 2013) *Command investigation into the grounding of USS Guardian (MCM 5) on Tubbataha Reef, Republic of the Philippines that occurred on 17 January 2013*. Tech Rep 5830 Ser N00/0639, Commander: United States Pacific Fleet, Department of the Navy, 250 Makalapa Drive, Pearl Harbor, Hawaii 96860-3131.
- Hankel, Lauren** (Nov 2012) *Assessment of the quality of OpenStreetMap data in Woodhill, Kameeldrift and Mamelodi for the use of location based services*. Bsc honours report, Department of Computer Science, University of Pretoria. Unpublished.
- Harber, Anton** (7 May 2015b) *Journalists have duty to verify social media news*. Business Day. URL <http://www.bdlive.co.za/opinion/columnists/2015/05/07/journalists-have-duty-to-verify-social-media-news>.
- Harrison, J A, Allan, D G, Underhill, L G, Herremans, M, Tree, A J, Parker, V & Brown, C J** (eds) (1997) *The Atlas of Southern African Birds; vol 1, Non-passerines, vol 2, Passerines*. BirdLife South Africa, Johannesburg.
- Harrison, J A, Underhill, L G & Barnard, P** (Mar – Apr 2008) *The seminal legacy of the Southern African Bird Atlas Project*. South African Journal of Science, **104**(3–4):pp 82–84.
- Harrison, Weber**. (23 Sep 2014) *PayPal merchants can now accept Bitcoin in North America, sort of*. VentureBeat. URL <http://venturebeat.com/2014/09/23/paypal-merchants-can-now-accept-bitcoin-in-north-america-sort-of/>.
- Harvey, Francis** (2009) *Commentary on Next Generation Digital Earth: More than Names — Digital Earth and/or Virtual Globes?* International Journal of Spatial Data Infrastructures Research, **4**:pp 111–116.
- Harvey, Francis** (2013a) *A new age of discovery: The post-GIS era*. In: **Jekel, T, Car, A, Strobl, J & Griesebner, G** (eds), *GI Forum 2013. Creating the GISociety*, Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/Offenbach, ISBN 978-3-87907-532-4, pp 272–281.
- Harvey, Francis** (2013b) *To volunteer or to contribute locational information? towards truth in labeling for crowdsourced geographic information*. In: Sui *et al* [2013], chap 3, pp 31–42.
- Harvey, Francis, Coetzee, Serena, Cooper, Antony K & Iwaniak, Adam** (23–28 Aug 2015) *Are the data sharing problems with sdis problems of zombies?* In: Robbi Sluter *et al* [2015b].
- Harvey, Francis, Iwaniak, Adam, Coetzee, Serena & Cooper, Antony K** (13–17 May 2012) *SDI Past, Present and Future: A Review and Status Assessment*. In: *Spatially Enabling Government, Industry and Citizens: Research and development perspectives*, chap 2, GSDI Association Press, Needham, MA, USA, pp 23–38. Presented at the GSDI 13 World Conference, Québec, Canada.
- Harvey, Francis, Jones, Jim, Scheider, Simon, Iwaniak, Adam, Kaczmarek, Iwona, Łukowicz, Jaromar & Strzelecki, Marek** (3–6 Jun 2014) *Little steps towards big goals. using linked data to develop next general sdpatial data infrastructures (aka SDI 3.0)*. In: **Huerta, Schade & Granell** (eds), *AGILE'2014 International Conference on Geographic Information Science*, Castellón, Spain.

## Bibliography

---

- Haslhofer, Bernhard & Isaac, Antoine** (21–23 Sep 2011) *data.europeana.eu — The European Linked Open Data Pilot*. In: *International Conference on Dublin Core and Metadata Applications 2011*, The Hague, The Netherlands, pp 94–104. URL <http://dcpapers.dublincore.org/pubs/article/view/3625>.
- Haunert, Jan-Henrik & Sester, Monika** (2008) *Assuring logical consistency and semantic accuracy in map generalization*. *Photogrammetrie-Fernerkundung-Geoinformation (PFG)*, 2008(3):pp 165–173.
- Hausmann, Natalie Suzette, McIntyre, Trevor, Bumby, Adam John & Loubser, Michael John** (Aug 2013) *Referencing practices in physical geography: How well do we cite what we write?* *Progress in Physical Geography*, 37(4):pp 543–549.
- Heald, Basil** (30 May 2011) *The value of collaboration and SDI services*. In: *Africa-GEO 2011: Spatial data infrastructure (SDI) workshop*, Committee for Spatial Information (CSI) and the Centre for Geoinformation Science, University of Pretoria, Cape Town. URL [http://web.up.ac.za/sitefiles/file/48/16053/Heald\\_2011\\_SDIWorkshop\\_TheValueOfCollaborationAndSDIServices\\_df.pdf](http://web.up.ac.za/sitefiles/file/48/16053/Heald_2011_SDIWorkshop_TheValueOfCollaborationAndSDIServices_df.pdf).
- Hebeler, John, Fisher, Matthew, Perez-Lopez, Andrew & Blace, Ryan** (2009) *Semantic Web Programming*. Wiley Publishing, Inc, Indianapolis, IN, ISBN 978-0-470-41801-7.
- Hellman, Eric** (19 Jan 2013a) *Edward Tufte was a Proto-Phreaker (#aaronswnyc Part 1)*. Go to Hellman. URL <http://go-to-hellman.blogspot.co.za/2013/01/edward-tufte-was-proto-phreaker>.
- Hellman, Eric** (25 Jan 2013b) *The four crimes of Aaron Swartz (#aaronswnyc part 2)*. Go to Hellman. URL <http://go-to-hellman.blogspot.co.za/2013/01/the-four-crimes-of-aaron-swartz>.
- Hendricks, Vincent F** (13 Aug 2014) *How social media is polarising us*. TechCentral. URL <http://www.techcentral.co.za/how-social-media-is-polarising-us/50310/>.
- Hern, Alex** (30 Oct 2013) *Darkmail opens: New email encryption standard aims to keep government agencies out*. The Guardian. URL <http://www.theguardian.com/technology/2013/oct/30/darkmail-encryption-inbox-silent-circle-lavabit>.
- Hernandez, Sergio, Wu, Huizhong, Daileida, Colin & Specia, Megan** (7 Jul 2015) *The value of the witness: How the London attacks changed the news*. Mashable. URL <http://mashable.com/2015/07/07/77-attacks-changed-the-news/>.
- Higgins, Eliot** (25 Jul 2015) *Searching the earth: Essential geolocation tools for verification*. Bellingcat. URL <https://www.bellingcat.com/resources/how-tos/2015/07/25/searching-the-earth-essential-geolocationtools-for-verification/>.
- Higgins, Eliot** “Brown Moses” (15 Jul 2014) *What is Bellingcat*. Brown Moses Blog. URL <http://brown-moses.blogspot.co.uk/>.
- Higgins, Sarah** (nd) *What are metadata standards*. Undated, URL <http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards>.

- Hjelmager**, Jan, **Delgado**, Tatiana, **Moellering**, Harold, **Cooper**, Antony K, **Danko**, David, **Huet**, Michel, **Aalders**, Henri J G L & **Martynenko**, Alexander (Jul 2005) *Developing a modelling for the spatial data infrastructure*. In: *22nd International Cartographic Conference*, A Coruña, Spain.
- Hjelmager**, Jan, **Moellering**, Harold, **Delgado**, Tatiana, **Cooper**, Antony K, **Rajabifard**, Abbas, **Rapant**, Petr, **Danko**, David, **Huet**, Michel, **Laurent**, Dominique, **Aalders**, Henri J G L, **Iwaniak**, Adam, **Abad**, Paloma, **Düren**, Ulrich & **Martynenko**, Alexander (2008) *An initial formal model for spatial data infrastructures*. *International Journal of Geographical Information Science*, **22**(11):pp 1295–1309.
- HLWIKI Canada** (2015a) *Metadata* — *HLWIKI Canada*. Tech rep, HLWIKI Canada. [Online; accessed 22-October-2015], URL <http://hlwiki.slais.ubc.ca/index.php?title=Metadata&oldid=141525>.
- HLWIKI Canada** (2015b) *Social cataloguing* — *HLWIKI Canada*. Tech rep, HLWIKI Canada. [Online; accessed 22-October-2015], URL [http://hlwiki.slais.ubc.ca/index.php?title=Social\\_cataloguing&oldid=141085](http://hlwiki.slais.ubc.ca/index.php?title=Social_cataloguing&oldid=141085).
- HLWIKI Canada** (2015c) *Social tagging* — *HLWIKI Canada*. Tech rep, HLWIKI Canada. [Online; accessed 22-October-2015], URL [http://hlwiki.slais.ubc.ca/index.php?title=Social\\_tagging&oldid=141021](http://hlwiki.slais.ubc.ca/index.php?title=Social_tagging&oldid=141021).
- Ho**, Chad (2015) *Get Chad to World Championships in Kazan, Russia!* Dreamfuel. URL <http://www.dreamfuel.me/athletes/206/campaigns/88>.
- Hobel**, Heidelinde & **Frank**, Andrew U (23 Sep 2014) *Exploiting linked spatial data and granularity transformations*. In: *Workshop on Geographic Information Observatories 2014 at the 8th International Conference on Geographic Information Science (GIScience 2014)*, Vienna, Austria, pp 15–22. Need to check the date and complete this reference, URL <http://ceur-ws.org/Vol-1273/paper2.pdf>.
- Hockey**, Phil A R, **Dean**, William Richard John & **Ryan**, Peter (2005) *Roberts' birds of southern Africa*. Trustees of the John Voelcker Bird Book Fund, Cape Town, seventh edn.
- Hodza**, Paddington (! Apr 2014) *Appreciative GIS and strength-based community change*. *Transactions in GIS*, **18**(2):pp 270–285.
- Hoenig**, Michael (9 Jun 2014a) *'unreliable' articles: More on peer review's frailties*. *New York Law Journal*. URL [http://www.herzfeld-rubin.com/publ\\_complexlitigation\\_20140609.htm](http://www.herzfeld-rubin.com/publ_complexlitigation_20140609.htm).
- Hoenig**, Michael (12 May 2014b) *'unreliable' articles, 'trial by literature' revisited*. *New York Law Journal*. URL [http://www.herzfeld-rubin.com/publ\\_complexlitigation\\_20140512.htm](http://www.herzfeld-rubin.com/publ_complexlitigation_20140512.htm).
- Hofman**, Elwin (2014) *An obligation of conscience: gossip as social control in an eighteenth-century Flemish town*. *European Review of History: Revue europ  enne d'histoire*, **21**(5):pp 653–670.

## Bibliography

---

- Holloway**, Adam, MP (2009) *The Failure of British Political and Military Leadership in Iraq*. Policy paper, First Defence. URL [http://www.firstdefence.org/Failure in Iraq.doc](http://www.firstdefence.org/Failure%20in%20Iraq.doc).
- Holputch**, Amanda (20 Sep 2013) *Brazil's controversial plan to extricate the internet from us control*. The Guardian. URL <http://www.theguardian.com/world/2013/sep/20/brazil-dilma-rousseff-internet-us-control>.
- Hopfstock**, Anja, **Buchroithner**, Manfred F & **Grünreich**, Dietmar (Nov 2013) *A conceptual framework for creating cartographic representations in SDI environments*. The Cartographic Journal, **50**(4):pp 345–355.
- Howe**, Jeff (2006) *The rise of crowdsourcing*. Wired Magazine, **14**(6).
- Huang**, Haosheng (23–28 Aug 2015) *Using social media data to study people's perception and knowledge of environments*. In: Robbi Sluter *et al* [2015b].
- Hudson**, Alex (5 Jun 2013) *Is small print in online contracts enforceable?* BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-22772321>.
- Hughes**, Simon (Dec 2005) *Geohydrology data model design: South African boreholes*. Master's thesis, University of Stellenbosch, Stellenbosch, South Africa.
- Hulac**, Benjamin (1 May 2015) *Tesla's Elon Musk unveils solar batteries for homes and small businesses*. Scientific American. URL <http://www.scientificamerican.com/article/tesla-s-elon-musk-unveils-solar-batteries-for-homes-and-small-businesses/>.
- Hume**, Tim & **Park**, Madison (1 Oct 2014) *Understanding the symbols of Hong Kong's 'Umbrella Revolution'*. CNN. URL <http://edition.cnn.com/2014/09/30/world/asia/objects-hong-kong-protest/index.html>.
- Ilves**, Toomas Hendrik (29 Sep 2013) *Keynote Address by President Toomas Hendrik Ilves at Panel Discussion "A Secure and Free Internet", the UN Dag Hammarskjöld Library Auditorium*. Permanent Mission of Estonia to the UN. URL [http://www.un.estemb.org/statements\\_articles/aid-930](http://www.un.estemb.org/statements_articles/aid-930).
- INCITS (21 May 2009) *ANSI/INCITS 453-2009, Information technology — North American Profile of ISO 19115:2003, Geographic information — Metadata (NAP — Metadata)*. International Committee for Information Technology Standards (INCITS), Washington DC, United States of America, 302 pp.
- Institute for Volunteering Research** (Jun 2010) *Using participatory mapping to explore participation in three communities*. Pathways through participation project report, Institute for Volunteering Research. URL <http://pathwaysthroughparticipation.org.uk/>.
- International Cartographic Association**, Commission III: Computer-assisted Cartography (1980) *Glossary of terms in Computer Assisted Cartography*. International Cartographic Association, 157 pp.
- International Statistical Institute (16–22 Aug 2009) *International Statistical Institute, 57th Biennial Session*, Durban, South Africa, ISBN 978-90-73592-29-2.

- Internet Hall of Fame** (2012) *Inductees Internet Hall of Fame Inventor: Tim Berners-Lee*. Internet Hall of Fame. URL <http://internethalloffame.org/inductees/tim-berners-lee>.
- Investors.com** (13 Jun 2013) *So Why Didn't NSA Catch The Tsarnaev Brothers?* Investor's Business Daily, Inc. URL <http://www.investors.com/politics/editorials/patriot-act-did-not-authorize-nsa-prism/>.
- Ioannidis, John P A** (2005) *Why most published research findings are false*. PLoS Medicine, 2(8):pp 0696–0701. E124.
- Irwin, Alan** (2001) *Constructing the scientific citizen: science and democracy in the biosciences*. Public Understanding of Science, 10:pp 1–18.
- Isaacson, David** (24 Jul 2015) *Cash-strapped swimmers Schoeman and Ho try crowd-funding*. Business Day. URL <http://www.bdlive.co.za/sport/othersport/2015/07/24/cash-strapped-swimmers-schoeman-and-ho-try-crowd-funding>.
- ISO 11620** (2008) *ISO 11620:2008, Information and documentation — Library performance indicators*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 15489-1** (2001) *ISO 15489-1:2001, Information and documentation — Records management — Part 1: General*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 15836** (2009) *ISO 15836:2009, Information and documentation — The Dublin Core metadata element set*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19101** (2002) *ISO 19101:2002, Geographic information — Reference model*. International Organization for Standardization (ISO), Geneva, Switzerland. Superseded.
- ISO 19101-1** (2014) *ISO 19101-1:2014, Geographic information — Reference model — Part 1: Fundamentals*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19113** (2002) *ISO 19113:2002, Geographic information — Quality principles*. International Organization for Standardization (ISO), Geneva, Switzerland. Withdrawn.
- ISO 19114** (2003) *ISO 19114:2003, Geographic information — Quality evaluation procedures*. International Organization for Standardization (ISO), Geneva, Switzerland. Withdrawn.
- ISO 19115** (2003) *ISO 19115:2003, Geographic information — Metadata*. International Organization for Standardization (ISO), Geneva, Switzerland. Superseded.
- ISO 19115-1** (1 Apr 2014) *ISO 19115:2014, Geographic information — Metadata — Part 1: Fundamentals*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19115-2** (2009) *ISO 19115-2:2009, Geographic information — Metadata — Part 2: Extensions for imagery and gridded data*. International Organization for Standardization (ISO), Geneva, Switzerland.

## Bibliography

---

- ISO 19115Cor1 (2006) *ISO 19115:2003, Geographic information — Metadata — Corrigendum 1*. International Organization for Standardization (ISO), Geneva, Switzerland. Superseded.
- ISO 19119 (2005) *ISO 19119:2005, Geographic information — Services*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19119Amd1 (2008) *ISO 19119:2005/Amd 1:2008, Geographic information — Services — Amendment 1*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19123 (2005) *ISO 19123:2005, Geographic information — Schema for coverage geometry and functions*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19128 (2005) *ISO 19128:2005, Geographic information — Web Map Server interface*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19131 (2007) *ISO 19131:2007, Geographic information — Data product specifications*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19136 (2007) *Geographic information — Geography Markup Language (GML)*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19138 (2006) *ISO/TS 19138:2006, Geographic information — Data quality measures*. International Organization for Standardization (ISO), Geneva, Switzerland. Withdrawn.
- ISO 19139 (2007) *ISO/TS 19139:2007, Geographic information — Metadata — XML schema implementation*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19139-1 (2014) *ISO/WD 19139-1, Geographic information — Metadata — XML schema implementation — Part 1*. International Organization for Standardization (ISO), Geneva, Switzerland. Working Draft.
- ISO 19139-2 (2012) *ISO/TS 19139-2:2012, Geographic information — Metadata — XML Schema Implementation — Part 2: Extensions for imagery and gridded data*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 19157 (2013) *ISO 19157:2013, Geographic information — Data quality*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 639-5 (2008) *ISO 639-5:2008, Codes for the representation of names of languages — Part 5: Alpha-3 code for language families and groups*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 639-6 (2009) *ISO 639-6:2009, Codes for the representation of names of languages — Part 6: Alpha-4 code for comprehensive coverage of language variants*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 8859-1 (1998) *ISO/IEC 8859-1:1998, Information technology — 8-bit single-byte coded graphic character sets — Part 1: Latin alphabet No. 1*. International Organization for Standardization (ISO), Geneva, Switzerland.

- ISO 9000 (2005) *ISO 9000:2005, Quality management systems — Fundamentals and vocabulary*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 9001 (2008) *ISO 9001:2008, Quality management systems — Requirements*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO 9004 (2009) *ISO 9004:2009, Managing for the sustained success of an organization — A quality management approach*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC (2004) *ISO/IEC Guide 2:2004, Standardization and related activities — General vocabulary*. International Organization for Standardization (ISO), Geneva, Switzerland, eighth edn.
- ISO/IEC (2011b) *ISO/IEC Directives, Part 2: Rules for the structure and drafting of International Standards*. International Organization for Standardization (ISO), Geneva, Switzerland, sixth edn.
- ISO/IEC 10646 (2011) *ISO/IEC 10646:2011, Information technology — Universal Coded Character Set (UCS)*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 10746-1 (1998) *ISO/IEC 10746-1:1998, Information technology — Open Distributed Processing — Reference Model: Overview*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 10746-2 (1996) *ISO/IEC 10746-2:1996, Information technology — Open Distributed Processing — Reference Model: Foundations*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 10746-3 (1996) *ISO/IEC 10746-3:1996, Information technology — Open Distributed Processing — Reference Model: Architecture*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 10746-4 (1998) *ISO/IEC 10746-4:1998, Information technology — Open Distributed Processing — Reference Model: Architectural Semantics*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 11179 (2003–2005) *ISO/IEC 11179, Information technology — Metadata registries (MDR)*. International Organization for Standardization (ISO), Geneva, Switzerland. In six parts, ISO/IEC 11179-1 [2004]; ISO/IEC 11179-2 [2005]; ISO/IEC 11179-3 [2003]; ISO/IEC 11179-4 [2004]; ISO/IEC 11179-5 [2005]; ISO/IEC 11179-6 [2005].
- ISO/IEC 11179-1 (2004) *ISO/IEC 11179-1:2004, Information technology — Metadata registries (MDR) — Part 1: Framework*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 11179-2 (2005) *ISO/IEC 11179-2:2005, Information technology — Metadata registries (MDR) — Part 2: Classification*. International Organization for Standardization (ISO), Geneva, Switzerland.

## Bibliography

---

- ISO/IEC 11179-3 (2003) *ISO/IEC 11179-3:2003, Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 11179-4 (2004) *ISO/IEC 11179-4:2004, Information technology — Metadata registries (MDR) — Part 4: Formulation of data definitions*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 11179-5 (2005) *ISO/IEC 11179-5:2005, Information technology — Metadata registries (MDR) — Part 5: Naming and identification principles*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 11179-6 (2005) *ISO/IEC 11179-6:2005, Information technology — Metadata registries (MDR) — Part 6: Registration*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 15444-2 (2004) *ISO/IEC 15444-2:2004, Information technology — JPEG 2000 image coding system — Part 2: Extensions*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 15938-5 (2003) *ISO/IEC 15938-5:2003, Information technology — Multimedia content description interface — Part 5: Multimedia description schemes*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 19501 (2005) *ISO/IEC 19501:2005, Information technology — Open Distributed Processing — Unified Modeling Language (UML) Version 1.4.2*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 19502 (2005) *ISO/IEC 19502:2005 Information technology — Meta Object Facility (MOF)*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 2382-17 (1999) *ISO/IEC 2382-17:1999, Information technology — Vocabulary — Part 17: Databases*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/IEC 8211 (1994) *ISO/IEC 8211:1994, Information technology — Specification for a data descriptive file for information interchange*. International Organization for Standardization (ISO), Geneva, Switzerland.
- ISO/TC 211 Ad hoc group on linked data** (20 Mar 2012) *Final report from the ad hoc group on linked data*. Tech Rep 211n3308, ISO/TC 211, Geographic information/Geomatics.
- ITU (2010) *The world in 2010: ICT facts and figures*. Tech rep, International Telecommunications Union (ITU), Geneva, Switzerland. URL <http://www.itu.int/ict>.
- ITU (2012) *Measuring the information society: 2012*. Tech rep, International Telecommunications Union (ITU), Geneva, Switzerland. URL [http://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2012/MIS2012\\_without\\_Annex\\_4.pdf](http://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2012/MIS2012_without_Annex_4.pdf).
- ITU (2013a) *Measuring the information society: 2013*. Tech rep, International Telecommunications Union (ITU), Geneva, Switzerland.

- ITU (Feb 2013b) *The world in 2013: ICT facts and figures*. Tech rep, International Telecommunications Union (ITU), Geneva, Switzerland. URL <http://www.itu.int/ict>.
- Jacqueline** “**Laika Spoetnik**” (8 May 2009) *Merck’s Ghostwriters, Haunted Papers and Fake Elsevier Journals*. Laika’s MedLibLog. Accessed 4 July 2014, URL <http://laikaspoetnik.wordpress.com/2009/05/08/mercks-ghostwriters-haunted-papers-and-fake-elsevier-journals/>.
- Jakes, Lara & Goldman, Adam** (15 Aug 2013) *Online chat rooms provide key havens for terrorist plotting*. Huff Post Tech. URL [http://www.huffingtonpost.com/2013/08/15/online-terrorist\\_n\\_3759444.html?view=print&comm\\_ref=false](http://www.huffingtonpost.com/2013/08/15/online-terrorist_n_3759444.html?view=print&comm_ref=false).
- James, Colin** (12 Sep 2014) *What Jennifer Lawrence should have done*. The New Zealand Herald. URL <http://www.nzherald.co.nz/news/print.cfm?objectid=11322727>.
- Jansen van Vuuren, Anna-Marie, Jansen van Vuuren, Joey & Venter, Suna** (5–6 Jul 2012) *The susceptibility of the South African Media to be used as a tool for information warfare*. In: *11th European Conference on Information Warfare and Security*, Laval, France, pp 127–134.
- Jiow, Hee Jhee Jiow & Lin, Julian** (7 Jan 2013) *The influence of parental factors on children’s receptiveness towards mobile phone location disclosure services*. First Monday, **18**(1), doi: 10.5210/fm.v18i1.4284. URL <http://firstmonday.org/ojs/index.php/fm/article/view/4284/3384>.
- Jobim, A C, De Moraes, V & Gimbel, N** (1962/1964) *The Girl from Ipanema* (Garota de Ipanema). Song lyrics.
- Johanis, Paul** (2005) *Documenting data elements in statistical agencies*. In: *Statistics Canada Symposium 2005: Methodological Challenges for Future Information Needs*, no 11-522-XIE in Statistics Canada International Symposium Series, Statistics Canada.
- Johnson, Thomas J, Bichard, Shannon L & Zhang, Weiwu** (2009) *Communication communities or “cyberghettos?”: A path analysis model examining factors that explain selective exposure to blogs*. Journal of Computer-Mediated Communication, **15**:pp 60–82.
- Kaczmarek, Iwona, Iwaniak, Adam & Łukowicz, Jaromar** (2014) *New spatial planning data access methods through the implementation of the INSPIRE Directive*. Real Estate Management and Valuation, **22**(1):pp 12–24.
- Kaherl, Amy** (12 Mar 2015) *How to Soup*. BBC News: Magazine. URL <http://www.bbc.com/news/magazine-31603493>.
- Kalantari, Mohsen, Olfat, Hamed & Rajabifard, Abbas** (2010) *Automatic spatial metadata enrichment: Reducing metadata creation burden through spatial folksonomies*. In: *GSDI-12*, Singapore.
- Kalantari, Mohsen, Rajabifard, Abbas, Olfat, Hamed & Williamson, Ian** (2014) *Geospatial metadata 2.0 — an approach for volunteered geographic information*. Computers, Environment and Urban Systems, **48**:pp 35–48.
- Kaminska, Izabella** “izakaminska2013” (31 Aug 2015) *Bitcoin zero-ville*. Dizzynomics:

## Bibliography

---

- Finding patterns in finance, econ and technology — probably where there are none. URL <https://dizzynomics.wordpress.com/2015/08/31/bitcoin-zero-ville/>.
- Kanakarajan**, Pavithra (8 May 2015) *Map maker will be temporarily unavailable for editing starting may 12, 2015*. Google Product Forums — Google Map Maker. URL <https://productforums.google.com/d/topic/map-maker/crFEbGXJ-HI>.
- Kaufman**, Charlie (2008) *Synecdoche, New York*. 124 minutes, colour.
- Kay**, John (18 Dec 2014) *Beware the supposed wisdom of crowds*. Business Day Live. URL <http://www.bdlive.co.za/opinion/2014/12/18/beware-the-supposed-wisdom-of-crowds>.
- Keall**, Chris (19 Jun 2014) *Slingshot launches Global Mode — allows access to overseas media*. The National Business Review. URL <http://www.nbr.co.nz/article/slingshot-launches-global-mode-%E2%80%93-allows-access-overseas-media-ck-141718>.
- Keeler**, Mary (2011) *Crowdsourced Knowledge: Peril and Promise for Conceptual Structures Research, Lecture Notes in Computer Science*, vol 6828/2011. Springer Verlag, pp 131–144, doi: 10.1007/978-3-642-22688-5\_10.
- Keen**, Andrew (2007) *The Cult of the Amateur: How Today's Internet is Killing Our Culture and Assaulting Our Economy*. Nicholas Brealey Publishing, London, ISBN 978-1-85788-393-0.
- Kelion**, Leo (22 Aug 2014) *NSA and GCHQ agents 'leak Tor bugs', alleges developer*. BBC News: Technology. URL <http://www.bbc.com/news/technology-28886462>.
- Kellaway**, Lucy (15 Dec 2014) *On work: Editorial decorum displayed as brown-nosing takes to Twitter*. Business Day Live, syndicated from The Financial Times. URL <http://www.bdlive.co.za/opinion/columnists/2014/12/15/on-work-editorial-decorum-displayed-as-brown-nosing-takes-to-twitter>.
- Kemp-Robertson**, Paul (16 Oct 2014) *Why your new bank will be Starbucks or Nike*. CNN. URL [http://edition.cnn.com/2014/08/05/opinion/starbucks-nike-currency/index.html?iid=article\\_sidebar](http://edition.cnn.com/2014/08/05/opinion/starbucks-nike-currency/index.html?iid=article_sidebar).
- Kennedy**, Christina (6 Jan 2015) *On the stage: The year of do-it-yourself theatre*. Business Day Live. URL <http://www.bdlive.co.za/life/entertainment/2015/01/06/on-the-stage-the-year-of-do-it-yourself-theatre>.
- Kermeliotis**, Teo (1 Aug 2014) *Making money: How to start your own currency*. CNN. URL <http://edition.cnn.com/2014/08/01/business/making-money-start-your-own-currency/index.html?hpt=bosread>.
- Kessler**, Glenn (4 Nov 2013) *A cautionary tale for politicians: Al Gore and the 'invention' of the Internet*. Washington Post. URL <https://www.washingtonpost.com/blogs/fact-checker/wp/2013/11/04/a-cautionary-tale-for-politicians-al-gore-and-the-invention-of-the-internet>.
- Kingsley**, Patrick (20 Jul 2011) *Avaaz: activism or 'slacktivism'?* The Guardian. URL

- <http://www.theguardian.com/world/2011/jul/20/avaaz-activism-slactivism-clicktivism>.
- Kirby, Dean** (7 Sep 2015) *Lake District hikers warned not to rely on mobile phones to plot their routes amid rise in rescue callouts*. The Independent. URL <http://www.independent.co.uk/news/uk/home-news/lake-district-hikers-warned-not-to-rely-on-mobile-phones-to-plot-their-routes-10488883>.
- Kizilkaya, Emre** (13 Jun 2013) *Behind Turkey's viral revolution, there are mad men (actually women)*. Huff Post World. URL [http://www.huffingtonpost.com/emre-kizilkaya/gezi-park-social-media\\_b\\_3435581.html](http://www.huffingtonpost.com/emre-kizilkaya/gezi-park-social-media_b_3435581.html).
- Klimushkin, Mikhail, Obiedkov, Sergei & Roth, Camille** (15–18 Mar 2010) *Approaches to the Selection of Relevant Concepts in the Case of Noisy Data*. In: *8th International Conference on Formal Concept Analysis*, Springer, Agadir, Morocco, pp 255–266.
- Kling, Rob** (2004) *The Internet and unrefereed scholarly publishing*. Annual review of information science and technology, **38**(1):pp 591–631.
- Kloppers, Nic** (Feb 2014) *Creating a national online GIS repository*. PositionIT. URL <http://www.ee.co.za/article/creating-national-online-gis-repository.html>.
- Knuth, Donald E** (1984) *The T<sub>E</sub>Xbook*. Addison-Wesley.
- Kokla, Margarita & Kavouras, Marinos** (2001) *Fusion of top-level and geographical domain ontologies based on context formation and complementarity*. International Journal of Geographical Information Science, **15**(7):pp 679–687, doi: 10.1080/13658810110061153.
- Kombuis, Koos** (2010) *Die Groen Fokkol Song*. URL <http://cdn.24.com/Cms/RelatedFiles/81/f089c13ad64a42889a894ef2df1e5b5b.mp3>.
- Kontopoulos, Stratos** (17 Mar 2015) *Monitoring the semantic drift of index terms by an evolving vector field: a novel experimental approach*. PERICLES Project Blog. URL [http://pericles-project.eu/blog/tag/semantic\\_drift](http://pericles-project.eu/blog/tag/semantic_drift).
- Kounadi, Ourania** (2009) *Assessing the quality of OpenStreetMap data*. Master's thesis, University College of London Department of Civil, Environmental And Geomatic Engineering, London, United Kingdom.
- Kourie, Derrick G & Oosthuizen, G Deon** (1998) *Lattices in machine learning: Complexity issues*. Acta Informatica, **35**:pp 269–292.
- Kramer, Robert** (1954) *Foreword: Literary and artistic products and copyright problems*. Law & Contemporary Problems, **19**(2):pp 139–140.
- Kravets, David** (3 Nov 2011) *Feds drop plan to lie in Public-Record Act requests*. Wired. URL <http://www.wired.com/threatlevel/tag/freedom-of-information-act/>.
- Kurtz, Lauren** (23 Jun 2015) *New Wyoming law criminalizes “unlawful collection of resource data”*. Climate Law Blog. URL <http://blogs.law.columbia.edu/climatechange/2015/06/23/new-wy-law-criminalizes-data-collection/>.

## Bibliography

---

- Kuznetsov**, Sergei O (6 Jun 2007) *On stability of a formal concept*. *Annals of Mathematics and Artificial Intelligence*, **49**(1–4):pp 101–115.
- Laber**, Terry, **Kish**, Paul, **Taylor**, Michael, **Owens**, Glynn, **Osborne**, Nikola & **Curran**, James (Jun 2014) *Reliability assessment of current methods in bloodstain pattern analysis*. Tech Rep 247180, National Institute of Justice, Washington, DC, USA. Award number 2010-DN-BX-K213. This document is a research report submitted to the US Department of Justice. This report has not been published by the Department.
- Lacy**, Sarah (17 Nov 2014) *The moment i learned just how far Uber will go to silence journalists and attack women*. Pando Media. URL <https://pando.com/2014/11/17/the-moment-i-learned-just-how-far-uber-will-go-to-silence-journalists-and-attack-women/>.
- Lambdin**, Charles (2012) *Significance tests as sorcery: Science is empirical—significance tests are not*. *Theory & Psychology*, **22**(1):pp 67–90. URL <http://tap.sagepub.com/content/22/1/67>.
- Lamport**, Leslie (1986) *TEX: A Document Preparation System*. Addison-Wesley.
- Lane**, Edwin (19 Aug 2014) *Google removes 12 BBC News links in ‘right to be forgotten’*. BBC News: Technology. URL <http://www.bbc.com/news/technology-28851366>.
- Lanier**, Jaron (30 May 2006) *Digital Maoism: The hazards of the new online collectivism*. Edge. URL [http://www.edge.org/3rd\\_culture/lanier06/lanier06\\_index.html](http://www.edge.org/3rd_culture/lanier06/lanier06_index.html).
- Lanier**, Jaron (27 Nov 2013a) *Digital passivity*. The New York Times. URL [www.nytimes.com/2013/11/28/opinion/digital-passivity](http://www.nytimes.com/2013/11/28/opinion/digital-passivity).
- Lanier**, Jaron (27 May 2013b) *Sell your data to save the economy and your future*. BBC News: Business. URL <http://www.bbc.co.uk/news/business-22658152>.
- Larivière**, Vincent & **Gingras**, Yves (2010) *On the prevalence and scientific impact of duplicate publications in different scientific fields (1980-2007)*. *Journal of Documentation*, **66**(2):pp 179–190.
- Lazer**, David, **Kennedy**, Ryan, **King**, Gary & **Vespignani**, Alessandro (2014) *The parable of Google Flu: Traps in big data analysis*. *Science*, **343**:pp 1203–1205.
- Le Clézio**, J M G (7 Dec 2008) *Nobel Lecture: In the forest of paradoxes*. Tech rep, The Nobel Foundation. Translated by Alison Anderson.
- Lee**, Dave (19 Apr 2013a) *Boston bombing: How internet detectives got it very wrong*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-22214511>.
- Lee**, Dave (25 Oct 2013b) *Wikipedia pilots articles-via-SMS service aimed at Africans*. BBC News: Technology. URL <http://www.bbc.co.uk/news/technology-24662267>.
- Lee**, Dave (21 Aug 2014a) *Diaspora social network cannot stop IS posts*. BBC News: Technology. URL <http://www.bbc.com/news/technology-28882042>.
- Lee**, Dave (28 Jul 2014d) *Police placing anti-piracy warning ads on illegal sites*. BBC News: Technology. URL <http://www.bbc.com/news/technology-28523738>.

- Lee, Timothy B** (1 Nov 2014e) *There are now 285 Bitcoin ATMs around the world*. Vox. URL <http://www.vox.com/2014/11/1/7139785/there-are-now-285-bitcoin-atms-around-the-world>.
- Lehrer, Jonah** (13 Dec 2010) *The truth wears off: Is there something wrong with the scientific method?* The New Yorker. URL <http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off>.
- Lemmen, C H J, van Oosterom, P J M, Uitermark, H T, Zevenbergen, J A & Cooper, A K** (28–30Sep 2011) *Interoperable Domain Models: The ISO Land Administration Domain Model LADM and its External Classes*. In: UDMS 2011, Delft, The Netherlands.
- Leonnig, Carol, Gold, Matea & Hamburger, Tom** (17 Sep 2013) *Military's background check system failed to block gunman with a history of arrests*. The Washington Post. URL [https://www.washingtonpost.com/politics/contractor-would-not-have-hired-aaron-alexis-if-past-brushes-with-law-had-been-known/2013/09/17/e5bc83da-1faa-11e3-8459-657e0c72fec8\\_story](https://www.washingtonpost.com/politics/contractor-would-not-have-hired-aaron-alexis-if-past-brushes-with-law-had-been-known/2013/09/17/e5bc83da-1faa-11e3-8459-657e0c72fec8_story).
- Lessig, Lawrence** (28 Dec 2005) *Creatives face a closed Net*. Financial Times. URL <http://www.ft.com/cms/s/2/d55dfe52-77d2-11da-9670-0000779e2340.html>.
- Levelt, W J M** (Commission Chair) (31 Oct 2011) *Interim-rapportage inzake door Prof dr D A Stapel gemaakte inbreuk op wetenschappelijke integriteit*. investigation report, Universiteit van Tilburg, Tilburg, The Netherlands. In Dutch.
- Levine, Dan** (1 Sep 2015) *Uber drivers' labor lawsuit granted class action status in California*. The Huffington Post. URL [http://www.huffingtonpost.com/entry/uber-drivers-labor-lawsuit-granted-class-action-status-in-california\\_us\\_55e60a54e4b0b7a9633aa657](http://www.huffingtonpost.com/entry/uber-drivers-labor-lawsuit-granted-class-action-status-in-california_us_55e60a54e4b0b7a9633aa657).
- Leyne, Jon** (11 Feb 2010) *The new cyber battlefield in Iran*. BBC News. URL [http://news.bbc.co.uk/go/pr/fr/-/2/hi/middle\\_east/8505645.stm](http://news.bbc.co.uk/go/pr/fr/-/2/hi/middle_east/8505645.stm).
- Li, Linna, F, Goodchild Michael & Xu, Bo** (2013) *Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr*. Cartography and Geographic Information Science, 40(2):pp 61–77.
- Library of Congress** (24 Apr 2008) *Marc to dublin core crosswalk*. Crosswalk, Library of Congress, Washington, DC, United States of America. URL <http://www.loc.gov/marc/marc2dc.html>.
- Lickliger, J C R** (23 Apr 1963) *Topics for discussion at the forthcoming meeting*. Memorandum for the members and affiliates of the Intergalactic Computer Network, Advanced Research Projects Agency, Washington, DC. Obtained via KurzweilAI.net, URL <http://www.kurzweilai.net/articles/art0366.html?printable=1>.
- Liebenberg, Elri** (Aug 2014) *Achieving the impossible: The 1:500 000 'Irrigation Map' of South Africa, 1935–1937*. The Cartographic Journal, 51(3):pp 237–248.
- Liebenberg, Louis** (Aug 2003) *A new environmental monitoring methodology*. Overview of CyberTracker, URL <http://www.cybertracker.org/>.

## Bibliography

---

- Liebenberg**, Louis, **Steventon**, Lindsay, **Benadie**, Karel & **Minye**, James (Jan–Dec 1999) *Rhino tracking in the Karoo National Park*. Pachyderm, (27).
- Lilien**, Gary L (Mar–Apr 2008) *The Ombudsman: Who's at Fault at Faulty Towers? Commentaries on the Citation Dilemma*. Interfaces, **38**(2):pp 123–124.
- Lisi**, Marco (Mar 2015) *A look into the future of positioning, navigation and timing: The world urgently needs an even stronger, more resilient and more versatile PNT infrastructure*. Coordinates : A resource on positioning, navigation and beyond. URL <http://mycoordinates.org/a-look-into-the-future-of-positioning-navigation-and-timing/>.
- Livingstone**, Rob (12 Jun 2013) *NSA leak could cause cloud repercussions: Pay careful attention to the nest of contractors and other providers running your cloud — if you can*. CFOcom. URL <http://ww2.cfo.com/data-security/2013/06/nsa-leak-could-cause-cloud-repercussions/>.
- Lloyd**, Matthew (3–5 Dec 2012) *Communicating with smartphones where there is no cellular infrastructure*. In: *United Nations International Expert Meeting on Crowdsourcing Mapping for Disaster Risk Management and Emergency Response*, Vienna, Austria. URL <http://www.un-spider.org/crowdsourcing-mapping>.
- Loftie-Eaton**, Megan (4 Jan 2015) *Comparing reporting rates between the First and Second Southern African Bird Atlas Project*. Ornithological Observations, **6**:pp 1–11. URL <http://oo.adu.org.za/content.php?id=163>.
- Longley**, D & **Shain**, M (1982) *Dictionary of Information Technology*. Macmillan Reference Books, 381 pp.
- Lopez-Pellicer**, Francisco J, **Vilches-Blázquez**, Luis M, **Zarazaga-Soria**, F Javier, **Muro-Medrano**, Pedro R & **Corcho**, Oscar (2011) *The Delft Report: Linked Data and the challenges for geographic information standardization*. Tech rep, Universidad de Zaragoza and Universidad Politécnica de Madrid, Spain. URL <http://oa.upm.es/9742/>.
- Loscalzo**, Joseph (13 Mar 2012) *Experimental irreproducibility: Causes, (mis)interpretations, and consequences*. Circulation, **125**(10):pp 1211–1214, doi: 10.1161/CIRCULATIONAHA.112.098244.
- Louw-Vaudran**, Liesl (14 May 2015) *Nigeria kidnappings: what a difference a hashtag makes*. ISS Today. URL <https://www.issafrica.org/iss-today/nigeria-kidnappings-what-a-difference-a-hashtag-makes>.
- Lugg**, Alexander (1 Jul 2013) *Mantous and alpacas as weapons of the weak: Chinese spoof video and self-expression online*. First Monday, **18**(7), doi: 10.5210/fm.v18i7.3885. URL <http://firstmonday.org/ojs/index.php/fm/article/view/3885/3695>.
- Lukhwareni**, T Joseph, **Madonsela**, Sibongile F, **Cooper**, Antony K, **Cronje**, Marius, **Mokhuwa**, Dineo E, **Podile**, Lucas M, **Pillay**, Nishan, **Maremba**, Thanyani & **Masemula**, Mandla (8 Jun 2005a) *ISO 19115 as the metadata standard for Statistics South Africa*. In: *ISO/TC 211 Workshop on Standards in Action*, Stockholm, Sweden.

- Lukhwareni**, T Joseph, **Madonsela**, Sibongile F, **Mokhuwa**, Dineo E & **Podile**, Lucas M (5–9 Sep 2005b) *Management of metadata in national statistical agency*. In: Commonwealth Statisticians [2005], p 6. URL <http://www.statssa.gov.za/commonwealth/speakerpresentations.asp>.
- Maake**, Moyagabo (7 Jan 2016) *Skywise's crowd-fund scheme shot down*. Business Day. URL <http://www.bdlive.co.za/business/transport/2016/01/07/skywises-crowd-fund-scheme-shot-down>.
- MacDowell**, Marsha (2012) *Project planning and management*. In: Boyd *et al* [2012]. URL <http://ohda.matrix.msu.edu/2012/06/project-planning-and-management/>.
- Maclean**, Gordon Lindsay & **Roberts**, Austin (1985) *Roberts' birds of southern Africa*. Trustees of the John Voelcker Bird Book Fund, Cape Town, sixth edn.
- MacManus**, Richard (17 Jan 2006) *Lessig on the read/write web*. ReadWriteWeb. URL [http://www.readwriteweb.com/archives/lessig\\_on\\_the\\_r.php](http://www.readwriteweb.com/archives/lessig_on_the_r.php).
- Makanga**, Prestige & **Smit**, Julian (29 Sep – 3 Oct 2008) *A review of the status of spatial data infrastructure implementation in Africa*. In: Coetzee *et al* [2008c]. URL <http://conference.osgeo.org/index.php/foss4g/2008/schedConf/presentations>.
- Makgoba**, M W (19 May 2000) *Editorial. hiv/aids: The peril of pseudoscience*. Science, 288(5469):p 1171, doi: 10.1126/science.288.5469.1171.
- Makinana**, Andisiwe (10 Sep 2013) *Zuma condemns 'opposite of the positive' SA media*. Mail & Guardian. URL <http://mg.co.za/print/2013-09-10-zuma-condemns-opposite-of-the-positive-sa-media>.
- Malakata**, Michael (12 Jun 2006) *Africa resolves telecommunications debate over EASSY project*. InfoWorld. URL <http://www.infoworld.com/t/communication-and-collaboration/africa-resolves-telecommunications-debate-over-eassy-project-891>.
- Malamud**, Randy (22 Dec 2013) *The new psychogeography of Tempelhof Airport, once a Nazi landmark*. The Atlantic. URL <http://www.theatlantic.com/technology/archive/2013/12/the-new-psychogeography-of-tempelhof-airport-once-a-nazi-landmark/282594/>.
- Malimabe**, Matile & **Jenneker**, Ashwell (4 Aug 2010) *Innovative technologies for statistical production: Experiences of Statistics South Africa*. Bulletin 33, IFC. The IFC's contribution to the 57th ISI Session, Durban, August 2009, URL [www.bis.org/ifc/publ/ifcb33f.pdf](http://www.bis.org/ifc/publ/ifcb33f.pdf).
- Mallaby**, Sebastian (28 Aug 2012) *American law is patent nonsense*. Financial Times. URL <http://www.ft.com/cms/s/2/ea9503c2-f0f9-11e1-89b2-00144feabdc0.html>.
- Mandiant** (2013) *APT1: Exposing One of China's Cyber Espionage Units*. Tech rep, Mandiant. URL <http://www.mandiant.com/apt1>.
- Mans**, Gerbrand (31 May–2 Jun 2011) *Developing a geo-data frame using dasymetric mapping principles to facilitate data integration*. In: Smit [2011]:p 16.

## Bibliography

---

- Mansourian, Ali & Abdolmajidi, Ehsan** (2011) *Investigating the system dynamics technique for the modeling and simulation of the development of spatial data infrastructures*. International Journal of Geographical Information Science, **25**(12):pp 2001–2023.
- Mäntylä, Mika V & Itkonen, Juha** (2013) *More testers — the effect of crowd size and time restriction in software testing*. Information and Software Technology, **55**:pp 986–1003.
- Marais, Stephan** (Mar 2010) *The definition and development of Open Innovation models to assist the innovation process*. Masters of engineering management, University of Stellenbosch, South Africa.
- Marcus, Griel** (23 Jul 1970) *Self Portrait No 25*. Rolling Stone.
- Marlinspike, Moxie** (12 Jun 2013) *We should all have something to hide*. Moxie Marlinspike Blog. URL <http://www.thoughtcrime.org/blog/we-should-all-have-something-to-hide/>.
- Marušić, Ana, Bošnjak, Lana & Jerončić, Ana** (2011) *A systematic review of research on the meaning, ethics and practices of authorship across scholarly disciplines*. PLoS ONE, **6**(9):p e23 477, doi: 10.1371/journal.pone.0023477.
- Marylebone Cricket Club** (1 Oct 2010) *The Laws of Cricket (2000 Code 4th Edition — 2010*. Marylebone Cricket Club (MCC), St John's Wood, London, United Kingdom. URL <http://www.lords.org/data/files/laws-of-cricket-2000-code-4th-edition-final-10422.pdf>.
- Maslow, Abraham H** (1943) *A theory of human motivation*. Psychological Review, **50**:pp 370–396, ISSN 1492-3713. Available from: “Classics in the History of Psychology: An internet resource developed by Christopher D. Green”, York University, Toronto, Ontario.
- Masnack, Mike** (14 Jun 2013) *Why the tech industry should be furious about NSA's over surveillance*. Techdirt. URL <https://www.techdirt.com/articles/20130614/12173323472/why-tech-industry-should-be-furious-about-nsas-over-surveillance>.
- Matthee, Karel** (9 Oct 2012) *Broadband for all: Connecting people from the ground up — one last mile at a time*. In: 4th CSIR Biennial Conference, Pretoria.
- Mazières, Antoine & Huron, Samuel** (24 May 2013) *Toward Google borders*. In: WebSci'13, Paris, France.
- McCann, Laurenellen** (5 Sep 2013) *Reasons (not) to release data*. Sunlight Foundation Blog. URL <http://sunlightfoundation.com/blog/2013/09/05/reasons-not-to-release-data/>.
- McCann, Laurenellen & Green, Alisha** (22 Oct 2013) *Empowering the open data dialogue*. Sunlight Foundation Blog. URL <http://sunlightfoundation.com/blog/2013/10/22/empowering-the-open-data-dialogue/>.
- McConchie, Alan** (2015) *Hacker cartography: Crowdsourced geography, OpenStreetMap, and the hacker political imaginary*. ACME: An International E-Journal for Critical Geographies, **14**(3):pp 874–898.

- McDougall**, Kevin (15–19 Jun 2009) *The potential of citizen volunteered spatial information for building SDI*. In: van Loenen *et al* [2009b].
- McDougall**, Kevin (11–16 Apr 2010) *From silos to networks — Will users drive spatial data infrastructures in the future?* In: FIG [2010].
- McKenzie**, Steven (10 Oct 2011) *Military jamming of GPS in Scotland suspended*. BBC News. Last updated at 15:20 GMT.
- McLaren**, Christine (18 Jan 2012) *New cartographers: How citizen mapmakers are changing the story of our lives*. Lab Notes. URL <http://blogs.guggenheim.org/lablog/the-new-cartographers-how-citizen-mapmakers-are-changing-the-story-of-our-lives/>.
- McLaren**, James (Sep 2009) *Why teens love MXit*. Parent24.com. The actual date of the blog is not given. Accessed 15 September 2009, URL [http://www.parent24.com/Teen\\_13-18/development\\_behaviour/Why-teens-love-MXit-20090911](http://www.parent24.com/Teen_13-18/development_behaviour/Why-teens-love-MXit-20090911).
- McLaren**, Robin (Nov 2011) *Crowdsourcing support of land administration: A new, collaborative partnership between citizens and land professionals*. Tech rep, Royal Institution of Chartered Surveyors.
- McLeod**, Duncan (28 Sep 2014) *Astounding growth in 3G in SA*. TechCentral. URL <http://www.techcentral.co.za/astounding-growth-in-3g-in-sa/51269/>.
- McMahon**, Brian (22–23 Aug 2015) *CODATA and (meta)data characterisation in the wider world*. In: *Metadata for raw data from X-ray diffraction and other structural techniques: A Satellite Workshop to the 29th European Crystallographic Meeting*, Rovinj, Croatia. Presentation slides.
- Meier**, Patrick (2012) *Crisis mapping in action: How open source software and global volunteer networks are changing the world, one map at a time*. Journal of Map & Geography Libraries: Advances in Geospatial Information, Collections & Archives, 8(2):pp 89–100.
- Meijer**, Ewout H, **Ben-Shakhar**, Gershon, **Verschuere**, Bruno & **Donchin**, Emanuel (Apr 2013) *A comment on Farwell (2012): brain fingerprinting: a comprehensive tutorial review of detection of concealed information with event-related brain potentials*. Cognitive Neurodynamics, 7(2):pp 155–158.
- Meo**, Sultan Ayoub (2014) *Open access journals: Open for rich, closed for poor*. Journal of the College of Physicians and Surgeons Pakistan, 24(8):p 611.
- Merritt**, Anna C, **Effron**, Daniel A & **Monin**, Benoît (2010) *Moral self-licensing: When being good frees us to be bad*. Social and Personality Psychology Compass, (4/5):pp 344–357.
- METIS** (7 Apr 2011) *Notes on the Generic Statistical Information Model (GSIM)*. Working Draft V0.1, Operationalize a Common Metadata/Information Management Framework (OCMIMF) Collaboration Team.
- Metz**, Cade (17 Jan 2013) *Facebook wants new breed of flash memory for storing old pics*. Wired. URL <http://www.wired.com/wiredenterprise/2013/01/facebook-cold-storage>.

## Bibliography

- Michael**, Chris (6 Oct 2014) *Missing Maps: nothing less than a human genome project for cities*. The Guardian. URL <http://www.theguardian.com/cities/2014/oct/06/missing-maps-human-genome-project-unmapped-cities>.
- Michalevsky**, Yan, **Boneh**, Dan, **Schulman**, Aaron & **Nakibly**, Gabi (2015) *PowerSpy: Location tracking using mobile device power analysis*. arXiv, (1502.03182v1).
- Microsoft News Center** (20 Nov 1997) *Now a Virtual Globe, Not Just a World Atlas*. Microsoft Press Release. Accessed 30 Jan 2010, URL <http://www.microsoft.com/presspass/press/1997/nov97/vglobepr.msp>.
- Mikkelsen**, Barbara (21 Jul 2014) *Sick child birthday card request*. snopescom: Rumor has it. URL <http://www.snopes.com/inboxer/medical/shergold.asp>.
- Minter**, Harriet (24 Sep 2014) *Hackers are trying to silence Emma Watson by leaking nude photos — but they only made her voice louder*. The Guardian. URL <http://www.theguardian.com/women-in-leadership/2014/sep/23/hackers-tried-silence-emma-watson-naked-photos-but-made-her-voice-louder>.
- Mitra**, S & **Rana**, V (2001) *Children and the Internet: Experiments with minimally invasive education in India*. The British Journal of Educational Technology, **32**(2):pp 221–232.
- Mixon**, J F & **Swyer**, W (2005) *Contribution, attribution and the assignment of intellectual property rights in economics*. Journal of Economic Studies, **32**:pp 382–386.
- Mkandawire**, Thandika (Sep 1997) *The social sciences in Africa: Breaking local barriers and negotiating international presence*. African Studies Review, **40**(2):pp 15–36. The Bashorun MKO Abiola Distinguished Lecture Presented to the 1996 African Studies Association Annual Meeting, URL <http://www.jstor.org/stable/525155>.
- Mkhize**, Nomalanga (8 Sep 2015) *Academia needs to instil worth of print*. Business Day. URL <http://www.bdlive.co.za/opinion/columnists/2015/09/08/academia-needs-to-instil-worth-of-print>.
- Moat**, Helen Susannah, **Curme**, Chester, **Avakian**, Adam, **Kenett**, Dror Y, **Stanley**, H Eugene & **Preis**, Tobias (8 May 2013) *Quantifying Wikipedia usage patterns before stock market moves*. Scientific Reports, **3**(1801):pp 1–5.
- Moellering** (Jan 1985) *Digital cartographic data standards: an interim proposed standard*. Tech Rep 6, National Committee for Digital Cartographic Data Standards.
- Moellering**, Harold (ed) (1991) *Spatial database transfer standards: current international status*. International Cartographic Association and Elsevier Applied Science, ISBN 1-85166-677-X.
- Moellering**, Harold (2000) *The scope and content of analytical cartography*. Cartography and Geographic Information Science, **27**(3):pp 205–223. Special Issue on Analytical Cartography.
- Moellering**, Harold, **Aalders**, Henri J G L & **Crane**, Aaron (eds) (2005) *World Spatial Metadata Standards: Scientific and Technical Characteristics, and Full Descriptions with Crosstable*. Elsevier and the International Cartographic Association, ISBN 0080439497.

- Moellering, Harold & Hogan, Richard** (eds) (1997) *Spatial database transfer standards 2: characteristics for assessing standards and full descriptions of the national and international standards in the world*. International Cartographic Association and Pergammon, ISBN 0-08-042433-3.
- Mohdin, Aamna** (17 Nov 2014) *Don't forget ethics when mapping uncharted slums*. SciDevNet. URL <http://www.scidev.net/global/health/scidev-net-at-large/ethics-mapping-uncharted-slums.html>.
- Moloto, Moloko** (7 Oct 2013) *Zuma invokes wrath of god*. IOL News. URL <http://www.iol.co.za/news/politics/zuma-invokes-wrath-of-god-1.1587841>.
- Monks, Kieron** (16 Oct 2014) *Sharing is daring: mapping the disruption economy*. CNN. URL [http://edition.cnn.com/2014/09/19/business/sharing-economy-guide/index.html?iid=article\\_sidebar](http://edition.cnn.com/2014/09/19/business/sharing-economy-guide/index.html?iid=article_sidebar).
- Moody, Glyn** (12 Jun 2013) *Is the US using Prism to engage in commercial espionage against Germany and others?* Techdirt. URL <https://www.techdirt.com/articles/20130611/10014923405/is-us-using-prism-to-engage-commercial-espionage-against-germany-others>.
- Mooney, Chris** (May/Jun 2011) *The science of why we don't believe science*. Mother Jones. URL <http://www.motherjones.com/politics/2011/03/denial-science-chris-mooney>.
- Mooney, P., Corcoran, P. & Winstanley, A.** (2010a) *A study of data representation of natural features in openstreetmap*. In: *Proceedings of GIScience*, p 150.
- Mooney, P., Corcoran, P. & Winstanley, A.C.** (2010b) *Towards quality metrics for openstreetmap*. In: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, pp 514–517.
- Mooney, Peter & Corcoran, Padraig** (2013) *Has OpenStreetMap a role in Digital Earth applications?* International Journal of Digital Earth, doi: 10.1080/17538947.2013.781688. In press.
- Mooney, Peter & Corcoran, Padraig** (2014) *Analysis of interaction and co-editing patterns amongst OpenStreetMap contributors*. Transactions in GIS, **18**(5):pp 633–659.
- Morita, Takashi** (23–28 Aug 2015) *Evolution of concepts in ubiquitous mapping*. In: Robbi Sluter *et al* [2015b].
- Morkel, T, Eloff, J H P & Olivier, M S** (Jun/Jul 2005) *An overview of image steganography*. In: *Proceedings of the Fifth Annual Information Security South Africa Conference (ISSA2005)*, Sandton, South Africa.
- Morris, Steven** (3 Sep 2009) *Made in Gwent with £10,000: the road safety video taking YouTube by storm*. The Guardian. URL <http://www.guardian.co.uk/uk/2009/sep/03/gwent-road-safety-film>.
- Morrison, R P** (Oct 2011) *Editorial: Retracted science and the retraction index*. Infection and Immunity, **79**(10):pp 3855–3859.

## Bibliography

---

- Moskvitch**, Katia (15 Jun 2012) *Ethiopia clamps down on Skype and other internet use on Tor*. BBC News: Technology. URL <http://www.bbc.com/news/technology-18461292>.
- Moylan**, Brian (23 Apr 2015) *A decade of YouTube has changed the future of television*. TIME. URL <http://time.com/3828217/youtube-decade/>.
- Mudau**, Nale (Oct 2010) *Spot building count supports informed decisions*. PositionIT.
- Muehlenhaus**, Ian (25–30 Aug 2013) *Four rhetorical styles of persuasive geocommunication: An initial taxonomy*. In: Buchroithner [2013].
- Müller-Maguhn**, Andy, **Poitrass**, Laura, **Rosenbach**, Marcel, **Sontheimer**, Michael & **Grothoff**, Christian (14 Sep 2014) *Map of the stars: The NSA and GCHQ Campaign Against German Satellite Companies*. The Intercept. URL <https://firstlook.org/theintercept/2014/09/14/nsa-stellar/>.
- Mullin**, Joe (10 Jan 2013b) *Why the UN's push to control the Internet isn't over*. Ars Technica. URL <http://arstechnica.com/tech-policy/2013/01/why-the-uns-push-to-control-the-internet-isnt-over/>.
- Mulupi**, Dinfan (27 Jan 2011) *Putting Kibera on the Map*. AudienceScapes: The InterMedia Knowledge Center. URL <http://www.audiencescapes.org/putting-kibera-map-kenya-OpenStreetMap-AMREF-Plan-International>.
- Murillo**, Angela P, **Thompson**, Cheryl A, **Carver**, Nico, **Robertson**, W Davenport, **Greenberg**, Jane & **Anderson**, William L (28–31 Oct 2012) *The data-at-risk initiative: Analyzing the current state of endangered scientific data*. In: ASIST 2012, Baltimore, MD, USA.
- Murray-Rust**, Peter (6 Dec 2013) *Can we trust commercial publishers or are we moving to 1984-like "publishers of truth"?* petermr's blog: A Scientist and the Web. URL <https://blogs.ch.cam.ac.uk/pmr/2013/12/06/can-we-trust-commercial-publishers-or-are-we-moving-to-1984-like-publishers-of-truth-we-must-act-now/>.
- Musk**, Elon (12 Jun 2014) *All our patent are belong to you*. Tesla Motors Blog. URL <http://www.teslamotors.com/blog/all-our-patent-are-belong-you>.
- MyBroadband (18 Oct 2010) *Broadband pricing: who gives the best value for money?* MyBroadband. URL <http://mybroadband.co.za/news/broadband/15907-Broadband-pricing-who-gives-the-best-value-for-money.html>.
- Nagpal**, Chirag & **Singhal**, Khushboo (31 Jul 2014) *Twitter user classification using ambient metadata*. arXiv, (arXiv:1407.8499v1 [cs.SI]).
- Narayanan**, Arvind (28 Jul 2011) *There is no such thing as anonymous online tracking*. The Center for Internet and Society at Stanford Law School. URL <http://cyberlaw.stanford.edu/node/6701>.
- Naroditskiy**, Victor, **Jennings**, Nicholas R, **Van Hentenryck**, Pascal & **Cebrian**, Manuel (2014) *Crowdsourcing contest dilemma*. Journal of The Royal Society Interface, 11(99), doi: 10.1098/rsif.2014.0532.
- National Academy of Sciences** (1990) *Spatial Data Needs: The Future of the National Mapping Program*. The National Academies Press. Mapping Science Committee; Commis-

- sion on Physical Sciences, Mathematics, and Resources; National Research Council, URL [www.nap.edu/catalog/9616.html](http://www.nap.edu/catalog/9616.html).
- National Research Council** (2007) *A Research Agenda for Geographic Information Science at the United States Geological Survey*. Tech rep, Committee on Research Priorities for the USGS Center of Excellence for Geospatial Information Science, Board on Earth Sciences and Resources (BESR), National Research Council of The National Academies. URL <http://www.nap.edu/catalog/12004.html>.
- National Research Council** (Apr 2015b) *Preparing the Workforce for Digital Curation*. The National Academies Press, ISBN 978-0-309-29694-6, 104 pp. Committee on Future Career Opportunities and Educational Requirements for Digital Curation; Board on Research Data and Information; Policy and Global Affairs; National Research Council, URL [http://www.nap.edu/catalog.php?record\\_id=18590](http://www.nap.edu/catalog.php?record_id=18590).
- Nature** (23 Oct 2012) *Shock and law: The Italian system's contempt for its scientists is made plain by the guilty verdict in L'Aquila (Editorial)*. *Nature*, **490**:p 446, doi: 10.1038/490446b.
- Naudé, Andries H, Van Huyssteen, Elsona, Maritz, Johan & Badenhorst, Willem** (17–18 Nov 2008) *Geospatial Analysis Platform: Supporting strategic spatial analysis and planning*. In: *2nd CSIR Biennial Conference: Science real and relevant*, Pretoria.
- Nebert, Douglas, Whiteside, Arliss & Vretanos, Panagiotis (Peter)** (eds) (23 Feb 2007a) *OpenGIS® Catalogue Services Specification*. 2.0.2, Open Geospatial Consortium Inc, Massachusetts, USA. OGC 07-006r1.
- Nebert, Douglas D** (2004) *Developing spatial data infrastructures: The SDI Cookbook*. Tech rep, Global Spatial Data Infrastructure Association (GSDI). URL <http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf>.
- Newman, Greg, Zimmerman, Don, Crall, Alycia, Laituri, Melinda, Graham, Jim & Stapel, Linda** (2010) *User-friendly web mapping: lessons from a citizen science website*. *International Journal of Geographical Information Science*, **24**(12):pp 1851–1869.
- Newman, Nathan** (17 Jul 2015) *Uber: When big data threatens local democracy*. The Huffington Post. URL [http://www.huffingtonpost.com/nathan-newman/uber-when-big-data-threat\\_b\\_7818452](http://www.huffingtonpost.com/nathan-newman/uber-when-big-data-threat_b_7818452).
- Nigeria, Federation of** (nd) *Criminal Code Act*. Nigeria, Federation of. URL <http://www.nigeria-law.org/Criminal Code Act-Part VI to the end.htm>.
- Nighswander, Tyler, Ledvina, Brent, Diamond, Jonathan, Brumley, Robert & Brumley, David** (16–18 Oct 2012) *GPS software attacks*. In: *CCS'12*, Raleigh, NC, USA.
- Nimmer, Melville B** (1954) *The right of publicity*. *Law & Contemporary Problems*, **19**(2):pp 203–223.
- Nkwinti, Gugile** (21 May 2010) *Notice 411 of 2010*. *Government Gazette*, (33205):pp 3–4. Appointment of members of the Committee for Spatial Information, by the Minister of Rural Development and Land Reform.

## Bibliography

---

- Norton**, Michael (2013) *Science by (social) media*. Edge. URL <https://edge.org/print/response-detail/23849>.
- Nothias**, Jean-Christophe (13 Jun 2013) *And now the second battle of the Internet*. Huff Post World. URL [http://www.huffingtonpost.com/jeanchristophe-nothias/internet-governance\\_b\\_3435812](http://www.huffingtonpost.com/jeanchristophe-nothias/internet-governance_b_3435812).
- NSA-Überwachung** (26 Oct 2013) *Merkels handy steht seit 2002 auf us-abhörliste*. Der Spiegel. URL <http://www.spiegel.de/politik/deutschland/nsa-ueberwachung-merkel-steht-seit-2002-auf-us-abhoerlistea-930193.html>.
- Nyhan**, Brendan (29 Sep 2014) *Why rumors outrace the truth online*. The New York Times. URL <http://www.nytimes.com/2014/09/30/upshot/its-so-much-more-fun-to-spread-rumors-than-the-truth.html>.
- Nyhan**, Brendan & **Reifler**, Jason (Oct 2013) *The effects of fact-checking threat: Results from a field experiment in the states*. Research paper, New America Foundation, Washington, DC, USA.
- Obiedkov**, Sergei, **Kourie**, Derrick G & **Eloff**, J H P (Feb–Mar 2009) *Building access control models with attribute exploration*. Computers and Security, **28**(1–2):pp 2–7.
- O’Brien**, Danny (8 Jan 2016) *Your apps, please? China shows how surveillance leads to intimidation and software censorship*. Electronic Frontier Foundation. URL <https://www.eff.org/deeplinks/2016/01/china-shows-how-backdoors-lead-software-censorship>.
- Odendaal**, Natasha (19 Feb 2014) *AfriForum, DWA clash over water quality tests*. Engineering News. URL <http://www.engineeringnews.co.za/article/afriforum-dwa-clash-over-water-quality-tests-2014-02-19>.
- Okoli**, Chitu, **Mehdi**, Mohamad, **Mesgari**, Mostafa, **Nielsen**, Finn Årup & **Lanamäki**, Arto (24 Oct 2012) *The people’s encyclopedia under the gaze of sages: A systematic review of scholarly research on Wikipedia*. Working paper, Concordia University, doi: 10.2139/ssrn.2021326. Available at SSRN, URL <http://ssrn.com/abstract=2021326> or <http://dx.doi.org/10.2139/ssrn.2021326>.
- Olfat**, Hamed, **Kalantari**, Mohsen, **Rajabifard**, Abbas & **Williamson**, Ian (2012) *Towards a foundation for spatial metadata automation*. Journal of Spatial Science, **57**(1):pp 65–81.
- Oliveira**, Italo Lopes & **Lisboa Filho**, Jugurta (27–30 Apr 2015) *A spatial data infrastructure review: Sorting the actors and policies from enterprise viewpoint*. In: 17th International Conference on Enterprise Information Systems (ICEIS), Barcelona, Spain. URL <http://www.dpi.ufv.br/~jugurta/papers/Oliveira%20-%20ICEIS2015.pdf>.
- Olivier**, Johan J, **Greenwood**, Peter H, **Cooper**, Antony K, **McPherson**, David R & **Engelbrecht**, Rudolph (Nov 1990) *Selecting a GIS for a national water management authority*. Photogrammetric Engineering & Remote Sensing, **56**(11):pp 1471–1475.
- O’Neill**, Mark (8 Aug 2012) *TunnelBear VPN circumvents geoblocking*. PCWorld. URL [http://www.pcworld.com/article/260236/tunnelbear\\_vpn\\_circumvents\\_geoblocking](http://www.pcworld.com/article/260236/tunnelbear_vpn_circumvents_geoblocking).

- Onsrud, Harlan & Rajabifard, Abbas** (eds) (Nov 2013) *Spatial Enablement in Support of Economic Development and Poverty Reduction: Research, Development and Education Perspectives*. GSDI Asociation Press, ISBN 978-0-9852444-2-2.
- O'Reilly, Tim** (30 Sep 2005) *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. O'Reilly Network. URL <http://oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1>.
- O'Reilly, Tim & Battelle, John** (2009) *Web Squared: Web 2.0 Five Years On*. Web<sup>2</sup> special report, O'Reilly Media, Inc.
- Ormeling, Ferjan** (Oct 2011) *Cartography and Geoinformation in the 20th and 21st Centuries*. meta-carto-semiotics: Journal for Theoretical Cartography, 4, ISSN 1868–1387. In conversation with Alexander Wolodtschenko and Florian Hruby.
- Orszag, Peter** (21 May 2009) *Democratizing data*. Office of Social Innovation and Civic Participation, The White House. URL <http://www.whitehouse.gov/blog/Democratizing-Data>.
- Orwell, George** (Summer 1946) *Why I write*. Gangrel, (4). URL <http://www.netcharles.com/orwell/essays/why-i-write.htm>.
- Orwell, George** (1949) *Nineteen eighty-four*. eBook No 0100021, Project Gutenberg of Australia. Date most recently updated: November 2008, URL <http://gutenberg.net.au/ebooks01/0100021.txt>.
- Ota, Morishige & Plews, Reese** (2015) *Development of a software tool as an introduction to geospatial information technology based on geospatial standards*. Cartography and Geographic Information Science, doi: 10.1080/15230406.2015.1031701.
- Owusu-Banahene, Wiafe, Mensah, Foster, Coetzee, Serena, Cooper, Antony K, Rautenbach, Victoria, Sinvula, Kisco M, Nangolo, Emma & Hippondoka, Martin** (Nov 2013) *A Description of Spatial Data Infrastructure Stakeholders in Ghana Using the ICA Model*, chap 4. In: Onsrud & Rajabifard [2013], pp 63–84.
- Oxford English Dictionary Department** (1973) *The Shorter Oxford English Dictionary on historical principles*. Clarendon Press, Oxford, 3rd edn. Reprinted with corrections, 1977.
- Oxford English Dictionary Department** (1981) *The Oxford dictionary for writers and editors*. Oxford University Press, Oxford. Reprinted with corrections, 1984.
- Pagliery, Jose** (11 Mar 2015) *The evolution of hacking*. CNNMoneycom. URL <http://edition.cnn.com/2015/03/11/tech/computer-hacking-history>.
- Palmer, Jason** (15 Jan 2010) *Tech tools offer Haiti lifeline*. BBC News. Accessed 30 Jan 2010, URL <http://news.bbc.co.uk/go/pr/fr/-/2/hi/technology/8461240.stm>.
- Panchaud, Nadia H, Iosifescu Enescu, Ionuț & Hurni, Lorenz** (23–28 Aug 2015) *Sharing cartographic knowledge with the crowd: on the complexity of cartographic rules*. In: Robbi Sluter et al [2015b].
- Pandor, Naledi** (24 Oct 2010) *Opening address by Minister Naledi Pandor MP*. In: 22nd International CODATA Conference, Spier, South Africa.

## Bibliography

---

- Pariser, Eli** (2012) *Beware online “filter bubbles”*. TED. Video with a transcript, URL [http://www.ted.com/talks/eli\\_pariser\\_beware\\_online\\_filter\\_bubbles.html](http://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles.html).
- Pariser, Eli** (7 May 2015) *Did Facebook’s big new study kill my filter bubble thesis?* Backchannel, Medium. URL <https://medium.com/backchannel/facebook-published-a-big-new-study-on-the-filter-bubble-here-s-what-it-says-ef31a292da95#.syl5yeu5w>.
- Parnas, David Lorge** (Nov 2007) *Stop the numbers game: Counting papers slows the rate of scientific progress*. Communications of the ACM, **50**(11):pp 19–21.
- Parsons, Ed** (25 Mar 2013) *Inspire a moonshot not a blueprint?* edparsonscom. URL <http://www.edparsons.com/>.
- Patashnik, Oren** (8 Feb 1988) *BIB<sub>T</sub>E<sub>X</sub>ing. Documentation for general BIB<sub>T</sub>E<sub>X</sub> users*.
- Peacock, Faansie** (2012) *Chamberlain’s LBJs: The Definitive Guide to Southern Africa’s Little Brown Jobs*. Mirafr Publishing, Pretoria.
- Pearson, Barbara, Randolph, Junius, Mosemak, Jerry & Durando, Jessica** (25 Feb 2015) *6 youtubers making more money than you*. USA Today. URL <http://www.usatoday.com/story/tech/2015/02/24/youtube-smosh-pewdiepie-jennamarbles-nigahiga-macbarbie07/23947527/>.
- Pejovic, Veljko, Johnson, David L, Zheleva, Mariya, Belding, Elizabeth M & Lysko, Albert** (7–10 Jan 2014) *VillageLink: Wide-area wireless coverage*. In: *6th International Conference on Communications Systems and Networks (COMSNETS 2014)*, Bangalore, India.
- Peled, Ammatzia & Cooper, Antony K** (23 Jul 2004) *Concepts of incremental updating and versioning: Current status*. In: **Altan, M Orhan** (ed), *XXth International Congress for Photogrammetry and Remote Sensing*, Istanbul, Turkey.
- Penny, Laurie** (22 Nov 2013) *Young and extremely clever digital activists are wasted in jail*. Mail & Guardian. URL <http://mg.co.za/article/2013-11-22-young-and-extremely-clever-digital-activists-are-wasted-in-jail>.
- Perez, Sarah** (11 May 2015) *Google shuts down Map Maker following hacks*. TechCrunch. URL <http://techcrunch.com/2015/05/11/google-shuts-down-map-maker-following-hacks>.
- Perkins, Chris** (8 Feb 2013) *Plotting practices and politics: (im)mutable narratives in OpenStreetMap*. Transactions of the Institute of British Geographers, doi: 10.1111/tran.12022.
- Peston, Robert** (2 Jul 2014) *Why has Google cast me into oblivion?* BBC News: Business. URL <http://www.bbc.com/news/business-28130581>.
- Peters, Elaine** (3 Nov 2015) *State, tourism and Uber curb car sales*. Business Day. URL <http://www.bdlive.co.za/economy/2015/11/03/state-tourism-and-uber-curb-car-sales>.
- Peterson, Elaine** (Nov 2006) *Beneath the metadata: Some philosophical problems with*

- folksonomy*. D-Lib Magazine, 12(11). URL <http://www.dlib.org/dlib/november06/peterson/11peterson.html>.
- Peuquet, D J** (1983) *A hybrid structure for the storage and manipulation of very large spatial data sets*. Computer Vision, Graphics, and Image Processing, 24:pp 14–27.
- Pfuhl, Gerit & Biegler, Robert** (2011) *Do humans know the imprecision inherent in a map? meta-carto-semiotics: Journal for Theoretical Cartography*, 4, ISSN 1868-1387.
- Phillipson, Gavin** (14 Jun 2013) *Q&A: The right to privacy*. BBC Religion & Ethics. URL <http://www.bbc.co.uk/religion/0/22887499?print=true>.
- Piatti, Barbara, Reuschel, Anne-Kathrin & Hurni, Lorenz** (Nov 2009) *Literary geography — or how cartographers open up a new dimension for literary studies*. In: 24th International Cartographic Conference, Santiago, Chile. URL [http://www.literaturatlas.eu/files/2012/02/Piatti2007\\_ICC\\_Chile.pdf](http://www.literaturatlas.eu/files/2012/02/Piatti2007_ICC_Chile.pdf).
- Pidot, Justin** (11 May 2015) *Forbidden data: Wyoming just criminalized citizen science*. Slate. URL [http://www.slate.com/articles/health\\_and\\_science/science/2015/05/wyoming\\_law\\_against\\_data\\_collection\\_protecting\\_ranchers\\_by\\_ignoring\\_the](http://www.slate.com/articles/health_and_science/science/2015/05/wyoming_law_against_data_collection_protecting_ranchers_by_ignoring_the).
- Pienaar, Heila & van Deventer, Martie** (eds) (11–13 May 2009) *African Digital Scholarship and Curation Conference 2009*, Pretoria, South Africa.
- Pierre, John M** (4 Feb 2001) *On the automated classification of web sites*. Linköping Electronic Articles in Computer and Information Science, 6(0). URL <http://www.ep.liu.se/ea/cis/2001/000/>.
- Pilkington, Ed** (14 Apr 2014) *Guardian and Washington Post win Pulitzer prize for NSA revelations*. The Guardian. URL <http://www.theguardian.com/media/2014/apr/14/guardian-washington-post-pulitzer-nsa-revelations>.
- Poore, Barbara S & Wolf, Eric B** (2013) *Metadata Squared: Enhancing Its Usability for Volunteered Geographic Information and the GeoWeb*, chap 4. In: Sui et al [2013], pp 43–64.
- Postmodernism Generator, The** (22 Sep 2012) *The Narrative of Meaninglessness: Marxist socialism and realism*. Communications From Elsewhere. A fake article created by The Posmodernism Generator, with the author nominally being a fictitious Anna F la Fournier., URL <http://www.elsewhere.org/pomo/>.
- Poyntz, Nick** (2009) *‘To have no neues is good neues’: licensed newsbooks during the early Commonwealth, 1649-1650*. Master of arts in comparative history of early modern societies, Birkbeck College, University of London.
- Poyntz, Nick** (13 Dec 2011) *Seventeenth-century crowd funding*. Mercurius Politicus. URL <https://mercuriuspoliticus.wordpress.com/2011/12/13/seventeenth-century-crowd-funding/>.
- PPGIS.net** (20 Jun 2015) *About PGIS*. PPGISnet. URL <http://www.ppgis.net/about-pgis/>.
- Priem, Jason** (28 Mar 2013) *Beyond the paper*. Nature, 495:pp 437–440.

## Bibliography

---

- Priss, Uta** (2006) *Formal concept analysis in information science*. Annual review of information science and technology, **40**:pp 521–543.
- Privat, Ludovic** (14 Apr 2009) *Egypt lifts ban on GPS*. GPS Business News. URL [http://www.gpsbusinessnews.com/Egypt-lifts-ban-on-GPS\\_a1463.html](http://www.gpsbusinessnews.com/Egypt-lifts-ban-on-GPS_a1463.html).
- Putra, Tandang Yuliadi Dwi** (Feb 2010) *A local spatial data infrastructure to support the Merapi volcanic risk management: A case study at Sleman Regency Indonesia*. PhD thesis, Universitas Gadjah Mada, Yogyakarta, Indonesia.
- Rajabifard, A, Feeney, Mary-Ellen F & Williamson, Ian P** (2002) *Future directions for SDI development*. International Journal of Applied Earth Observation and Geoinformation, **4**(1):pp 11–22.
- Rajabifard, Abbas & Coleman, David** (eds) (2012) *Spatially Enabling Government, Industry and Citizens: Research and Development Perspectives*. GSDI Association Press, Québec, Canada, ISBN 978-0-9852444-2-2.
- Rak, Andriy** (Apr 2013) *Legal issues and validation of volunteered geographic information*. Master's thesis, Department of Geodesy and Geomatics Engineering, University of New Brunswick, Fredericton, NB, Canada.
- Rakitienskaia, Anastassia, Olivier, Martin S & Cooper, Antony K** (15–17 Aug 2011) *Nature and forensic investigation of crime in Second Life*. In: 10th Annual ISSA Conference, Johannesburg, South Africa.
- Rakitienskaia, Anastassia Sergeevna** (2015) *Digital Forensics in Second Life*. Master's thesis, University of Pretoria, Pretoria, South Africa. URL <http://repository.up.ac.za/handle/2263/44263>.
- Rambaldi, Giacomo** (30 Aug 2013) *At global land rights conference, combining participatory mapping tools with traditional knowledge emerges as powerful weapon to fight massive land grabs*. PPgisnet Blog. URL <http://participatorygis.blogspot.co.za/2013/08/at-global-land-rights-conference>.
- Ramirez, Edith** (6 Jan 2015) *Privacy and the IoT: Navigating policy issues*. In: International Consumer Electronics Show, Las Vegas, Nevada. Opening Remarks of FTC Chairwoman Edith Ramirez.
- Ramsey, Paul** (27 Sep 2006) *Why sdis fail*. Clever elephant: a perspicacious pachyderm. URL <http://blog.cleverelephant.ca/2006/09/why-sdis-fail.html>.
- Ranganathan, Shiyali Ramamrita** (1931) *The Five Laws of Library Science*. No 2 in Publication Series, Madras Library Association, Madras, India.
- Ranganathan, Shiyali Ramamrita** (1951) *Classification and Communication*. University of Delhi, Delhi, India. URL <http://hdl.handle.net/10150/105279>.
- Raper, Jonathan, Rhind, David & Shepherd, John** (1992) *Postcodes: the new geography*. Longman Scientific & Technical.
- Rautenbach, Victoria** (31 May– 2 Jun 2011) *Fundamental spatial datasets for municipalities*. In: Smit [2011].

- Rautenbach, Victoria & Coetzee, Serena** (25–30 Aug 2013) *Books for SDI education and training in South Africa*. In: Buchroithner [2013].
- Rautenbach, Victoria, Coetzee, Serena & Iwaniak, Adam** (2012a) *Orchestrating OGC web services to produce thematic maps in a spatial information infrastructure*. Computers, Environment and Urban Systems. In press.
- Rautenbach, Victoria, Coetzee, Serena, Smit, Julian, du Plessis, Heindrich & Muzondo, Ivan Farayi** (2–4 Oct 2012b) *Identifying the target audiences, media and messages for SDI education and training in South Africa*. In: Coetzee [2012].
- Rautenbach, Victoria-Justine** (18 Jan 2013) *Orchestrating standard web services to produce thematic maps in a geoportal of a spatial data infrastructure*. Master's thesis, University of Pretoria, Pretoria, South Africa.
- Rens, Andrew** (15 Jun 2007) *What is the role of iCommons? ex Africa semper aliquid novi*. Accessed 4 July 2014, URL <http://aliquidnovi.org/what-is-the-role-of-icommons/>.
- Reporters Without Borders** (2015) *World Press Freedom Index 2015: "National Security" — Spurious Grounds*. Reporters Without Borders. URL <http://index.rsrf.org/#!/themes/national-security-spurious-grounds>.
- Retief, Ernst** (Jun 2011) *Is it a bird?* BirdLife South Africa E-newsletter:p 4.
- Reuteurs** (25 Nov 2012) *Africa chokes telecoms growth*. Fin24. URL <http://www.fin24.com/Economy/Africa-chokes-telecoms-growth-20121125>.
- Rey-Moreno, Carlos, Tucker, William D, Bidwell, Nicola J, Roro, Zukile, Siya, Masbulele Jay & Simo-Reigadas, Javier** (11–12 Jan 2013) *Experiences, challenges and lessons from rolling out a rural WiFi mesh network*. In: DEV '13, ACM, Bangalore, India. URL 978-1-4503-1856-3.
- Reynolds, Emma** (12 Sep 2014) *Zilla van den Born faked Asia trip on Facebook while still in her Amsterdam flat*. newscomau. URL <http://www.news.com.au/technology/online/social/zilla-van-den-born-faked-asia-trip-on-facebook-while-still-in-her-amsterdam-flat/news-story/0df93efb564076c2319ed39222b2aa5f>.
- Rhind, David & Openshaw, Stan** (29 Mar– 3 Apr 1987) *The BBC Domesday System: A nation-wide GIS for \$4448*. In: AutoCarto [1987], pp 595–603. URL <http://www.mapcontext.com/autocarto/proceedings/auto-carto-8/>.
- Rifkin, Jeremy** (7 Apr 2014) *The internet of things: Monopoly capitalism vs. collaborative commons*. Huffington Post. URL [http://www.huffingtonpost.com/jeremy-rifkin/internet-of-things\\_b\\_5104072](http://www.huffingtonpost.com/jeremy-rifkin/internet-of-things_b_5104072).
- Rigaux, Philippe, Scholl, Michel & Voisard, Agnès** (2002) *Spatial databases with application to GIS*. Morgan Kaufmann Publishers, San Francisco.
- Rivest, Ronald L** (1 Jul 1998) *Chaffing and winnowing: Confidentiality without encryption*. Tech rep, MIT Lab for Computer Science. URL <http://people.csail.mit.edu/rivest/chaffing-980701.txt>.

## Bibliography

- Robb**, Alice (18 Sep 2014) *The duck penis paradox: Is too much Internet pop science drowning out the serious stuff?* New Republic. URL <https://newrepublic.com/article/119404/duck-penis-paradox-plos-ones-mission-democratize-science>.
- Robbi Sluter**, Claudia, **Madureira Cruz**, C B & **Leal de Menezes**, P M (eds) (23–28 Aug 2015a) *Cartography — Maps connecting the world — 27th International Cartographic Conference (ICC 2015)*. Springer, Rio de Janeiro, Brazil, ISBN 978-3-319-17737-3, 386 pp. URL <http://www.springer.com/978-3-319-17737-3>.
- Robbi Sluter**, Claudia, **Madureira Cruz**, Carla Bernadete, **Camboim**, Silvana Philippi, **Delazari**, Luciene Stamato, **do Couto Fernandes**, Manoel, **Silva de Barros**, Rafael, **Firkowski**, Henrique & **Leal de Menezes**, Paulo Márcio (eds) (23–28 Aug 2015b) *27th International Cartographic Conference (ICC 2015)*, Rio de Janeiro, Brazil.
- Robertson**, A, **Simmons**, R E, **Jarvis**, A M & **Brown**, C J (1995) *Can bird atlas data be used to estimate population size? a case study using Namibian endemics*. Biological Conservation, **71**(1):pp 87–95.
- Robertson**, M P, **Cumming**, G S & **Erasmus**, B F N (2010) *Getting the most out of atlas data*. Diversity and Distributions, **16**:pp 363–375.
- Robinson**, Peter (5 Jan 2015) *Madonna: How the control queen lost her touch when media went social*. The Guardian. URL <http://www.theguardian.com/music/musicblog/2015/jan/05/madonna-control-queen-social-media-rebel-heart>.
- Robson**, Edward S (Sep 2012) *Responding to liability: Evaluating and reducing tort liability for digital volunteers*. Policy series, The Woodrow Wilson Center's Commons Lab.
- Roever**, Carrie L, **van Aarde**, R J & **Leggett**, K (2013) *Functional connectivity within conservation networks: Delineating corridors for African elephants*. Biological Conservation, **157**:pp 128–135.
- Romer**, Paul M (May 2015) *Mathiness in the theory of economic growth*. American Economic Review: Papers & Proceedings 2015, **105**(5):pp 89–93.
- Ronan**, Paul, **Poffenberger**, Michael & **Geyer**, Chelsea (Apr 2013) *Hidden in plain sight: Sudan's Harboring of the LRA in the Kafia Kingi Enclave, 2009-2013*. Tech rep, The Resolve LRA Crisis Initiative, with the Enough Project and Invisible Children.
- Roos**, Adrian (2015) *GIS and the road ahead*. PositionIT. URL <http://www.ee.co.za/article/gis-road-ahead.html>.
- Ropeik**, David (22 Oct 2012) *The L'Aquila verdict: A judgment not against science, but against a failure of science communication*. Scientific American. URL <http://blogs.scientificamerican.com/guest-blog/2012/10/22/the-laquila-verdict-a-judgment-not-against-science-but-against-a-failure-of-sciencecommunication/>.
- Roper**, Chris (1 Aug 2013) *Zim elections: Ignorance is not bliss*. Mail & Guardian. URL <http://mg.co.za/print/2013-08-01-zim-elections-ignorance-is-not-bliss>.

- Ross, Elliot** (Jul 2013) *How drug companies' PR tactics skew the presentation of medical research*. The Guardian. URL <http://www.theguardian.com/science/2011/may/20/drug-companies-ghost-writing-journalism>.
- Roswell, Charles** (ed) (1 Jun 2009) *ISO/TC 211 Standards Guide*. ISO/TC 211, *Geographic information/Geomatics*.
- Roth, Camille, Obiedkov, Sergei & Kourie, Derrick G** (30 Oct–1 Nov 2006) *Towards concise representation for taxonomies of epistemic communities*. In: Ben Yahia & Mephu Nguifo [2006], pp 205–218.
- Roth, Camille, Obiedkov, Sergei & Kourie, Derrick G** (2008a) *On succinct representation of knowledge community taxonomies with formal concept analysis*. International Journal of Foundations of Computer Science, **19**(2):pp 383–404.
- Rout, Milanda** (9 Apr 2009) *Doctors signed Merck's Vioxx studies*. The Australian. URL <http://www.theaustralian.news.com.au/business/story/0,,25311725-36418,00.html>.
- Rushe, Dominic** (3 Oct 2013) *Lavabit founder refused fbi order to hand over email encryption keys*. The Guardian. URL <http://www.theguardian.com/world/2013/oct/03/lavabit-ladar-levison-fbi-encryption-keys-snowden>.
- Russia Today** (3 Jul 2013) *'Restore the Fourth': Reddit, Mozilla, thousands of people set for July 4 NSA spying protest*. RT: Question More. URL <http://rt.com/usa/nsa-online-protest-rally-611>.
- Rutkin, Aviva Hope** (14 Aug 2013) *"spoofers" use fake GPS signals to knock a yacht off course*. MIT Technology Review, (v1.13.05.10). URL <http://www.technologyreview.com/news/517686/spoofers-use-fake-gps-signals-to-knock-a-yacht-off-course/>.
- Salmon, Felix** (3 Apr 2013) *The Bitcoin Bubble and the Future of Currency*. Money & Banking. URL <https://medium.com/@felixsalmon/the-bitcoin-bubble-and-the-future-of-currency-2b5ef79482cb#.eg7c6w40z>.
- Sample, Ian** (9 Dec 2013) *Nobel winner declares boycott of top science journals*. The Guardian. URL <http://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals>.
- Samuels, Crystal, Brown, Qunita, Leoschut, Lezanne, Jantjies, Janine & Burton, Patrick** (2013) *Connected dot com. young people's navigation of online risks: Social media, icts & online safety*. Tech rep, Centre for Justice and Crime Prevention (CJCP) and UNICEF South Africa. Part of the 2012 National School Violence Study (NSVS).
- Sanchez, Ray** (27 Feb 2015) *Prominent Bangladeshi-American blogger Avijit Roy killed*. CNNcom. URL <http://edition.cnn.com/2015/02/27/asia/bangladeshi-american-blogger-dead/>.
- Sánchez-Vaquerizo, Javier Argota, Alcocer, Atxu Amann y & Gutiérrez, Rodrigo Delso** (2015) *Hacking, reverse engineering and mapping of public open data through collaborative platforms*. Plurimond: An International Forum for Research and Debate on Human

## Bibliography

---

- Settlements, 7(15). URL <http://web.poliba.it/plurimondi/index.php/Plurimondi/article/view/232>.
- SANS 1876 (2013) *SANS 1876, Feature instance identification standard (draft)*. South African Bureau of Standards (SABS), Pretoria, South Africa.
- SANS 1878 (2005) *SANS 1878-1:2005, South African spatial metadata standard, Part 1 — Core metadata profile*. South African Bureau of Standards (SABS), Pretoria, South Africa.
- SAPA (11 Jul 2013) *Baba Jukwa, 'Zimbabwe's own Julian Assange'*. Voices of Africa. URL <http://voicesofafrica.co.za/baba-jukwa-zimbabwes-own-julian-assange/>.
- Saxton**, Gregory D, **Oh**, Onook & **Kishore**, Rajiv (2013) *Rules of crowdsourcing: Models, issues, and systems of control*. Information Systems Management, 30(1):pp 2–20.
- Scheepers**, C Ferdi (2–7 Apr 1989) *Vector-based computer graphics in automated map compilation*. In: *Auto-Carto IX: Proceedings of the International Symposium on Computer-Assisted Cartography*, Baltimore, MD, USA, pp 724–734. URL <http://mapcontext.com/autocarto/proceedings/auto-carto-9/index.html>.
- Scheepers**, C Ferdi, **van Biljon**, Willem R & **Cooper**, Antony K (Sep 1986) *Guidelines to set up a classification for geographical information*. Internal Report I723, NRIMS CSIR.
- Schiermeier**, Quirin (10 Jul 2012) *I was sued for libel under an unjust law*. Nature, 487(141), doi: 10.1038/487141a. URL [www.nature.com/news/i-was-sued-for-libel-under-an-unjust-law-1.10979](http://www.nature.com/news/i-was-sued-for-libel-under-an-unjust-law-1.10979).
- Schmitz**, Peter M U & **Cooper**, Antony K (11–13 May 2009) *Process flows for data archiving in a government department*. In: Pienaar & van Deventer [2009].
- Schmitz**, Peter M U & **Cooper**, Antony K (11–14 Jul 2011) *Using cellular telephones to track participants' movements to and from an event*. In: *South African Transportation Conference (SATC)*, Pretoria.
- Schmitz**, Peter M U, **Cooper**, Antony K, **Byleveld**, Piet & **Rossmo**, D Kim (9–12 Dec 2000) *Using GIS and digital aerial photography to assist in the conviction of a serial killer*. In: *4th Annual International Crime Mapping Research Conference*, San Diego, California, USA. URL <http://researchspace.csir.co.za/dspace/handle/10204/2778>.
- Schmitz**, Peter M U, **Cooper**, Antony K, **Kruger**, Tinus, **Speed**, Kenneth, **Barkhuizen**, Michael, **Lochner**, Hennie & **Linnen**, Chris (23–28 Aug 2015) *Space-Time Visualization for Investigative and Forensic Purposes*. In: Robbi Sluter et al [2015a], pp 267–281. URL <http://www.springer.com/978-3-319-17737-3>.
- Schneiderman**, Eric T (23 Sep 2013) *Attorney General Schneiderman announces agreement with 19 companies to stop writing fake online reviews and pay more than \$350,000 in fines*. New York State Office of the Attorney General. URL <http://www.ag.ny.gov/press-release/ag-schneiderman-announces-agreement-19-companies-stop-writing-fake-online-reviews-and>.
- Schneier**, Bruce (4 Oct 2013) *Attacking Tor: how the NSA targets users' online*

- anonymity*. The Guardian. URL <http://www.theguardian.com/world/2013/oct/04/tor-attacks-nsa-users-online-anonymity>.
- Schneier:2014** (4 Dec 2014) *Why Uber's 'god view' is creepy*. CNN. URL <http://edition.cnn.com/2014/12/04/opinion/schneier-uber-privacy-issue/index>.
- Schofield**, Hugh (27 Jun 2012) *Minitel: The rise and fall of the France-wide web*. BBC News Magazine. URL <http://www.bbc.co.uk/news/magazine-18610692>.
- Schuurman**, Nadine Cato (Apr 2000b) *Critical GIS: theorizing an emerging science*. PhD thesis, The University of British Columbia, Vancouver, Canada.
- Schwabe**, Craig A & **Govender**, Sives (Mar 2010a) *Getting geoinformation and SDI to work for Africa — Part 1*. PositionIT:pp 41–46.
- Schwabe**, Craig A & **Govender**, Sives (Apr–May 2010b) *Getting geoinformation and SDI to work for Africa — Part 2*. PositionIT:pp 41–45.
- Schwabe**, Craig A & **Govender**, Sives (May 2012) *Stakeholder survey on defining the criteria and identifying core geospatial datasets and data custodians in South Africa*. Tech rep, AfricaScope and EIS-Africa. Prepared for the Development Bank of Southern Africa.
- Schwartz**, Lisa M, **Woloshin**, Steven & **Baczek**, Linda (5 Jun 2002) *Media coverage of scientific meetings: Too much, too soon?* Journal of the American Medical Association, 287(21):pp 2859–2863.
- Schweller**, Randall L (16 Jun 2014) *The age of entropy: Why the new world order won't be orderly*. Foreign Affairs. URL <https://www.foreignaffairs.com/articles/united-states/2014-06-16/age-entropy>.
- Scott**, Mark (4 Jul 2014b) *Google reinstates some links in Europe*. The New York Times. URL <http://www.nytimes.com/2014/07/05/business/international/google-to-guardian-forget-about-those-links-right-to-be-forgotten-bbc>.
- SDMX (2009) *SDMX content-oriented guidelines, Annex 4: Metadata common vocabulary*. Guideline, SDMX.
- Sebag-Montefiore**, Clarissa (12 Aug 2013) *Micro movies beat China's censors*. BBC Culture. URL <http://www.bbc.com/culture/story/20130812-micro-movies-beat-chinas-censors>.
- Seitz**, Justin (5 Oct 2015) *Gangs of Detroit: OSINT and indictment documents*. Bellingcat. URL <https://www.bellingcat.com/resources/2015/10/05/gangsof-detroit-osint-and-indictment-documents/>.
- Sekercioglu**, Cagan H. (2013) *Citation opportunity cost of the high impact factor obsession*. Current Biology, 23(17):pp R1–R2.
- Self**, Will (2007) *Psychogeography*. Bloomsbury, 255 pp. Illustrated by Ralph Steadman.
- Seton**, Maria, **Williams**, Simon, **Zahirovic**, Sabin & **Micklethwaite**, Steven (9 Apr 2013) *Obituary: Sandy Island (1876–2012)*. Eos, Transactions American Geophysical Union, 94(15):pp 141–148.

## Bibliography

---

- Sharman, Andy** (13 Oct 2014) *Listed companies get in on crowdfunding*. Business Day. URL <http://www.bdlive.co.za/life/2014/10/13/listed-companies-get-in-on-crowdfunding>.
- Shaw, Darren** (28 Nov 2013b) *Improve local rankings by removing spammers*. The Whitespark Blog. URL <http://www.whitespark.ca/blog/post/19-improve-localrankings-by-removingspammers>.
- Shein, Esther** (Sep 2013) *Ephemeral data*. Communications of the ACM, **56**(9):pp 20–22.
- Shen, Cenyu & Björk, Bo-Christer** (2015) ‘predatory’ open access: a longitudinal study of article volumes and market characteristics. BMC Medicine, **13**(230):pp 1–15, doi: 10.1186/s12916-015-0469-2.
- Shiels, Maggie** (26 Jun 2009) *Web slows after Jackson’s death*. BBC News. URL <http://news.bbc.co.uk/1/hi/technology/8120324.stm>.
- Shiels, Maggie** (14 Apr 2010) *Congress to archive every tweet*. BBC News. URL <http://news.bbc.co.uk/go/pr/fr/-/2/hi/technology/8621297.stm>.
- Shirky, Clay** (2005) *Ontology is Overrated: Categories, Links, and Tags*. Web page. Clay Shirky’s Writings About the Internet. Accessed 4 July 2012, URL [http://www.shirky.com/writings/ontology\\_overrated.html](http://www.shirky.com/writings/ontology_overrated.html).
- Shopes, Linda** (2012) *Making sense of oral history*. In: Boyd *et al* [2012]. URL <http://ohda.matrix.msu.edu/2012/08/making-sense-of-oral-history/>.
- Shurkin, Joel N** (19 Jul 2012) *Landsat looks and sees*. NASA. URL [http://www.nasa.gov/mission\\_pages/landsat/news/landsat-history.html](http://www.nasa.gov/mission_pages/landsat/news/landsat-history.html).
- Siddle, James** (10 Apr 2014) *I know where you were last summer: London’s public bike data is telling everyone where you’ve been*. The Variable Tree. URL <http://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer>.
- Siebritz, Lindy-Anne** (2014) *Assessing the accuracy of OpenStreetMap data in South Africa for the purpose of integrating it with authoritative data*. Master’s thesis, University of Cape Town. URL [open.uct.ac.za/handle/11427/9148](http://open.uct.ac.za/handle/11427/9148).
- Siebritz, Lindy-Anne & Fourie, Helena** (11–13 Aug 2015) *The South African Spatial Data Infrastructure: a collaborative SDI*. In: *Geomatics Indaba*, Ekurhuleni, South Africa.
- Siebritz, Lindy-Anne, Sithole, George & Zlatanova, S** (aug/sep 2012) *Assessment of the homogeneity of volunteered geographic information in South Africa*. In: XXII ISPRS Congress, Melbourne, Australia.
- Siegal, Jacob** (11 May 2015) *End of an era: You won’t see peeing Android mascots in Google Maps anymore*. BGR Media, LLC. URL <http://bgr.com/2015/05/11/google-maps-map-maker-shut-down/>.
- Silberbauer, Michael & Geldenhuys, Willie** (29 Sep – 3 Oct 2008) *Using Keyhole Markup Language to create a spatial interface to South African water resource data through Google Earth*. In: Coetzee *et al* [2008c]. URL <http://conference.osgeo.org/index.php/foss4g/2008/schedConf/presentations>.

- Silberzahn, Raphael & Uhlmann, Eric L** (8 Oct 2015) *Many hands make tight work*. *Nature*, **526**:pp 189–191.
- Sillito, David** (5 Apr 2013) *Libraries to store all uk web content*. BBC News: Entertainment and Arts. URL <http://www.bbc.co.uk/news/entertainment-arts-22028738>.
- Simkin, John** (Aug 2014) *Is Wikipedia under the control of political extremists?* Spartacus Blog. URL <http://spartacus-educational.com/spartacus-blogURL29.html>.
- Simonite, Tom** (9 Jul 2013) *Build your own Internet with mobile mesh networking*. MIT Technology Review, (v1.13.05.10). URL <http://www.technologyreview.com/news/516571/build-your-own-internet-with-mobile-mesh-networking/>.
- Sinvula, Kisco M, Coetzee, Serena, Cooper, Antony K & Hipondoka, Martin** (2–4 Oct 2012) *Exploring the potential suitability of an SDI model in context of the National Spatial Data Infrastructure (NSDI) of Namibia*. In: Coetzee [2012].
- Sinvula, Kisco M, Coetzee, Serena, Cooper, Antony K, Nangolo, Emma, Owusu-Banahene, Wiafe, Rautenbach, Victoria & Hipondoka, Martin** (25–30 Aug 2013) *A contextual ICA stakeholder model approach for the Namibian Spatial Data Infrastructure (NamSDI)*. In: Buchroithner *et al* [2013], pp 381–394, doi: 10.1007/978-3-642-32618-9.
- Slee, Tom** (Jun 2013) *Notes against openness*. FutureEverything. URL <http://futureeverything.org/2013/03/notes-against-openness>.
- Smale, Will** (13 Jan 2014a) *The crowdfunding boss inspired by her parents' struggles*. BBC News: Business. URL <http://www.bbc.co.uk/news/business-25619253>.
- Smale, Will** (30 Sep 2014b) *Taboola: The internet firm at the forefront of 'click-bait'*. BBC News: Business. URL <http://www.bbc.com/news/business-29322578>.
- Smillie, Shaun** (13 Oct 2014) *Journal 'fails the test'*. Times Live. URL <http://www.timeslive.co.za/thetimes/2014/10/13/journal-fails-the-test>.
- Smit, Julian** (ed) (31 May–2 Jun 2011) *AfricaGEO 2011*, Cape Town, South Africa.
- Smit, Julian, Makanga, Prestige, Lance, Kate & de Vries, Walter** (15–19 Jun 2009) *Exploring relationships between municipal and provincial government SDI implementers in South Africa*. In: van Loenen *et al* [2009b].
- Smith, Barry** (15 Sep 2010) *Ontology: the why and how explained*. presentation. Director of the National Center for Ontological Research at the State University of New York (SUNY) at Buffalo.
- Sobel, Dava** (1998) *Longitude: The true story of a lone genius who solved the greatest scientific problem of his time*. Fourth Estate, London.
- Sokal, Alan D** (Oct 1996a) *Transgressing the boundaries: An afterword*. *Philosophy and Literature*, **20**(2):pp 338–346.
- Sokal, Alan D** (Spring/Summer 1996b) *Transgressing the boundaries: Towards a transformative hermeneutics of quantum gravity*. *Social Text*, (46/47):pp 217–252.

## Bibliography

---

- Solove, Daniel J** (2007) *"i've got nothing to hide" and other misunderstandings of privacy*. San Diego Law Review, **44**. URL [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=998565](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=998565).
- South Africa** (1999) *Statistics Act (Act No 6 of 1999)*. G 19957, 21 April 1999, Government Printer.
- South Africa** (2000) *Promotion of Access to Information Act (Act No 2 of 2000)*. G 20852, 3 February 2000, Government Printer. URL <http://www.info.gov.za/view/DownloadFileAction?id=68186>.
- South Africa** (2003) *Spatial Data Infrastructure Act (Act No 54 of 2003)*. G 25973, 4 February 2004, Government Printer. URL <http://www.info.gov.za/view/DownloadFileAction?id=68032>.
- South Africa** (2008a) *Intellectual Property Rights from Publicly Financed Research and Development Act (Act No 51 of 2008)*. G 31745, 22 December 2008, Government Printer. URL <http://www.info.gov.za/view/DownloadFileAction?id=94343>.
- South Africa** (2010b) *Protection of Information Bill*. G 32999, 5 March 2010, Government Printer. URL <http://www.info.gov.za/>.
- South Africa** (2011) *South African Weather Service Amendment Bill*. G 34648, 30 September 2011, Government Printer. URL <http://www.info.gov.za/>.
- South Africa** (15 Aug 2012) *National Development Plan 2030: Our future — make it work*. National Planning Commission, ISBN 978-0-621-41180-5. URL <http://www.gov.za/documents/download.php?f=172306>.
- South Africa** (11 Nov 2013c) *Protection of State Information Bill*. B 6H—2010, Government Printer. URL <http://www.info.gov.za/>.
- South Africa** (5 Aug 2013d) *Spatial Planning and Land Use Management Act (Act No 16 of 2013)*. G 36730, Government Printer. URL <http://www.info.gov.za/>.
- Souza, Wagner D, Lisboa-Filho, Jugurta, Vidal Filho, Jarbas N & Câmara, Jean H S** (24–27 Nov 2013) *DM4VGI: A template with dynamic metadata for documenting and validating the quality of volunteered geographic information*. In: XIV GEOINFO, Campos do Jordão, Brazil.
- SparkNotes Editors** (nd) *Sparknote on discipline and punish*. SparkNotes LLC. Accessed 31 October 2013, URL <http://www.sparknotes.com/philosophy/disciplinepunish/>.
- Sprigman, Christopher Jon & Granick, Jennifer** (2013) *U.S. government surveillance: Bad for silicon valley, bad for democracy around the world*. The Atlantic. URL <http://www.theatlantic.com/technology/archive/2013/06/us-government-surveillance-bad-for-silicon-valley-bad-for-democracy-around-the-world/277335/>.
- Stallman, Richard M** (Jul 2008a) *The anatomy of a trivial patent*. GNU Operating System. Accessed 4 July 2014, URL <http://www.gnu.org/philosophy/trivial-patent.html>.

- Stallman**, Richard M (Oct 2008b) *Did you say “intellectual property”? it's a seductive mirage*. GNU Operating System. Accessed 4 July 2014, URL <http://www.gnu.org/philosophy/notipr.html>.
- Stallman**, Richard M (2014a) *Reasons not to use uber*. Richard Stallman's personal site. URL <https://stallman.org/uber.html>.
- Stamos**, Alex (12 Jan 2013) *The truth about Aaron Swartz's “crime”*. Unhandled Exception: Building Better Internets. URL <http://unhandled.com/2013/01/12/the-truth-about-aaron-swartzs-crime/>.
- Stark**, Hans-Jörg (2011) *Quality assessment of volunteered geographic information using open Web map services within OpenAddresses*. In: *Proceedings of the Geoinformatics Forum Salzburg*, Wichmann Heidelberg, pp 101–110. URL [http://gispoint.de/fileadmin/user\\_upload/paper\\_gis\\_open/537509015.pdf](http://gispoint.de/fileadmin/user_upload/paper_gis_open/537509015.pdf).
- Stark**, Hans-Jörg (2012) *A field report on the role of free and open source geospatial software at University of Applied Sciences*. In: *Open Source Geospatial Research & Education Symposium (OGRS)*, pp 75–79. URL [http://ogrs2012.heig-vd.ch/public/ogrs2012/abstracts/A\\_Field\\_Report\\_on\\_the\\_Role\\_of\\_Free\\_and\\_Open\\_Source\\_Geospatial\\_Software\\_at\\_the\\_University\\_of\\_Applied\\_Sciences.pdf](http://ogrs2012.heig-vd.ch/public/ogrs2012/abstracts/A_Field_Report_on_the_Role_of_Free_and_Open_Source_Geospatial_Software_at_the_University_of_Applied_Sciences.pdf).
- Statistics Canada (Oct 2003) *Statistics Canada Quality Guidelines*. Guideline 12-539-XIE, Statistics Canada. Downloaded on 5 October 2005, URL <http://www.statcan.ca/english/freepub/12-539-XIE/index.htm>.
- Statistics South Africa (2008) *South African Statistical Quality Assessment Framework*. Manual, Statistics South Africa, Pretoria, South Africa. URL <http://www.statssa.gov.za/nss/index.asp>.
- Steadman**, Ian (7 May 2013) *Wary of Bitcoin? A guide to some other cryptocurrencies*. Wired UK. URL <http://www.wired.co.uk/news/archive/2013-05/7/alternative-cryptocurrencies-guide/viewall>.
- Stein**, Jason (23 Jul 2013) *Wisconsin enacts legislation targeting food stamp trafficking*. Governing. URL <http://www.governing.com/news/state/mct-wi-enacts-legislation-targeting-food-stamp-trafficking>.
- Stevens**, S S (1946) *On the theory of scales of measurement*. *Science*, **103**:pp 677–680.
- Stirling**, Peter, **Chevallier**, Philippe & **Illien**, Gildas (2012) *Web archives for researchers: Representations, expectations and potential uses*. *D-Lib*, **18**(3/4).
- Stone**, Jon (24 Dec 2015) *Theresa May wants to see your internet history, so we thought it was only fair to ask for hers*. *The Independent*. URL <http://www.independent.co.uk/news/uk/politics/theresa-may-wants-to-see-your-internet-history-so-we-thought-it-was-only-fair-to-ask-for-hers-a6785591>.
- Stöver**, Cathrin (27–30 Oct 2002) *European outreach efforts*. In: *Fall 2002 Internet2 Member Meeting*, Los Angeles, CA, USA.

## Bibliography

---

- StrategyPage.com (2013) *Chinese GPS open for business*. StrategyPagecom. URL <https://www.strategypage.com/htm/htspace/20130114>.
- Sturges, Paul & Neill, Richard** (2004) *The Quiet Struggle: Information and Libraries for the People of Africa*. Omnie Interactive, ISBN 5 070050 078682 77.
- Sui, Daniel Z** (2008) *The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS*. Computers, Environment and Urban Systems, **32**(1):pp 1–5.
- Sui, Daniel Z, Elwood, Sarah & Goodchild, Michael F** (eds) (2013) *Crowdsourcing Geographic Knowledge*. Springer.
- Sullivan, Bob** (29 Jun 2013) *The 'Internet of Things' pits George Jetson vs. George Orwell*. NBC News. URL <http://www.nbcnews.com/technology/internet-things-pits-george-jetson-vs-george-orwell-6C10462818>.
- Sundgren, Bo** (1995) *Guidelines for the modelling of statistical data and metadata*. *Questiö*, **19**:pp 321–357.
- Supreme Court of the United States** (19 Mar 2001) *Ohio v Matthew Reiner*. Case No. 532, \_ U.S. \_ (2001) 1, Supreme Court of the United States, United States of America.
- Swain, Frank** (19 Jul 2013) *Can you spend a week without cash?* BBC Future. URL <http://www.bbc.com/future/story/20130719-how-i-spent-a-week-without-cash>.
- Swartz, Aaron** (2013) *Aaron Swartz's A Programmable Web: An Unfinished Work*. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, ISBN 9781627051699.
- Swartz, Aaron & Hendler, James** (Oct 2001) *The Semantic Web: A Network of Content for the Digital City*. In: *Proceedings Second Annual Digital Cities Workshop*, Kyoto, Japan. URL <http://blogspace.com/rdf/SwartzHendler>.
- Swartz, Robert, Swartz, Susan, Swartz, Noah, Swartz, Ben & Stinebrickner-Kauffman, Taren** (13 Jan 2013) *Remember Aaron Swartz: Official Statement from the family and partner of Aaron Swartz*. Tumblr.com. URL <http://gautamghosh.tumblr.com/post/40393663042/remember-aaron-swartz-official-statement-from-the>.
- Tabuchi, Hiroko** (2 Aug 2015) *Chinese textile mills are now hiring in places where cotton was king*. The New York Times. URL <http://nyti.ms/1LY8sIq>.
- Tadeo, Maria** (5 Nov 2014) *Alex from Target has teenage girls swooning — but was it a hoax? tech start-up claims internet sensation was a marketing experiment*. The Independent. URL <http://www.independent.co.uk/news/business/news/alex-from-target-has-teenage-girls-around-the-world-swooning-but-was-it-a-hoax-a-tech-start-up-9841283>.
- Takahashi, Dean** (23 Sep 2014) *Why Rudy Giuliani wants Noriega to get his bloody hands off Activision's Call of Duty profits*. GamesBeat. URL <http://venturebeat.com/2014/09/23/whyrudy-giuliani-wants-noriega-to-get-his-bloody-handsoff-activisions-call-of-duty-profits/>.

- Tan**, Geraldine (16 Sep 2010) *ISO award puts geographic information standardizers on the map*. Tech rep, International Organization for Standardization (ISO), Geneva, Switzerland. URL <http://www.iso.org/iso/pressrelease.htm?refid=Ref1354>.
- Tansley**, Scott (7 Jun 2014a) *Why SDI's fail — a 2014 perspective — Part 1*. 50 shades of geodata sharing. URL <http://www.50shades.net.nz/>.
- Tansley**, Scott (9 Jun 2014b) *Why SDI's fail — a 2014 perspective — Part 2*. 50 shades of geodata sharing. URL <http://www.50shades.net.nz/>.
- Tarboton**, Warwick (30 Jan 2011) *Nylsvley 2011 woodland census, 12th year*. Friends of Nylsvley and the Nyl Floodplain. URL <http://www.nylsvley.co.za/?go=census1>.
- Taub**, Amanda (24 Oct 2014) *Can digital cameras stop war criminals? this activist thinks so*. Vox. URL <http://www.vox.com/2014/10/24/7032287/ryan-boyette-q-a>.
- Tavakoli-Far**, Nastaran (7 Feb 2013) *Artists use data to make political statements*. BBC News: Magazine. URL <http://www.bbc.co.uk/news/magazine-21018205>.
- Taylor**, Matthew, **Hopkins**, Nick & **Kiss**, Jemima (1 Nov 2013) *NSA surveillance may cause breakup of internet, warn experts*. The Guardian. URL <http://www.theguardian.com/world/2013/nov/01/nsa-surveillance-cause-internet-breakup-edward-snowden>.
- Templeton**, Brad (30 Nov 1991) *Dear Emily Postnews*. Usenet newsgroup newsanswers. URL [http://w2.eff.org/Net\\_culture/Net\\_info/EFF\\_Net\\_Guide/EEGTTI\\_HTML/eeg\\_272.html](http://w2.eff.org/Net_culture/Net_info/EFF_Net_Guide/EEGTTI_HTML/eeg_272.html).
- Tett**, Gillian (5 Jan 2015) *How social media split the family*. Business Day Live, syndicated from The Financial Times Limited. URL <http://www.bdlive.co.za/opinion/columnists/2015/01/05/how-social-media-split-the-family>.
- Tharyan**, Prathap (Jan–Dec 2012) *Criminals in the citadel and deceit all along the watchtower: Irresponsibility, fraud, and complicity in the search for scientific truth*. Mens Sana Monographs, 10(1):pp 158–180, doi: 10.4103/0973-1229.91426.
- The Onion** (27 May 2015) *FIFA frantically announces 2015 Summer World Cup in United States: Global soccer tournament to kick off in America later this afternoon*. The Onion, 51(21). URL <http://www.theonion.com/article/fifa-frantically-announces-2015-summer-world-cup-u-50525>.
- The Royal Parks (nd) *Assemblies, Demonstrations, Rallies and Marches*. Park policy, The Royal Parks. URL [http://www.royalparks.org.uk/documents/main/docs/demonstrations\\_and\\_assemblies.pdf](http://www.royalparks.org.uk/documents/main/docs/demonstrations_and_assemblies.pdf).
- Theunissen**, E (Sep 2014) *So you think you are safe: About (mis)perceptions, unwarranted assumptions and unintended capabilities*. Coordinates: A resource on positioning, navigation and beyond. URL <http://mycoordinates.org/so-you-think-you-are-safe/>.
- Thompson**, Ken (Aug 1983) *Reflections on trusting trust*. Communications of the ACM, 27(8):pp 761–763.
- Thurston**, Jeff (12 Aug 2013) *Geospatial or geomatics: The headaches of terminology in Canada*.

## Bibliography

- GoGeomatics Canada Magazine. URL <http://www.gogeomatics.ca/magazine/geospatial-or-geomatics-the-headaches-of-terminology-in-canada.htm>.
- Thygesen, Lars & Giovannini, Enrico** (16–22 Aug 2009) *Wikis, dynamic charts, videos and other innovative tools to transform statistics into knowledge*. In: International Statistical Institute [2009].
- Timberg, Craig** (24 Aug 2014) *For sale: systems that can secretly track where cellphone users go around the globe*. Washington Post. URL [https://www.washingtonpost.com/business/technology/for-sale-systems-that-can-secretly-track-where-cellphone-users-go-around-the-globe/2014/08/24/f0700e8a-f003-11e3-bf76-447a5df6411f\\_story](https://www.washingtonpost.com/business/technology/for-sale-systems-that-can-secretly-track-where-cellphone-users-go-around-the-globe/2014/08/24/f0700e8a-f003-11e3-bf76-447a5df6411f_story).
- Tomlinson, Roger F** (1988) *The impact of the transition from analogue to digital cartographic representation*. The American Cartographer, **15**(3):pp 249–261.
- Topping, Alexandra** (31 May 2015) *Ex-Fifa vice president Jack Warner swallows Onion spoof*. The Guardian. URL <http://www.theguardian.com/football/2015/may/31/ex-fifa-vice-president-jack-warner-swallows-onion-spoof>.
- Tracks4Africa** (2011) *Botswana: Traveller's map*. Map, Tracks4Africa Enterprises (Pty) Ltd, Technopark, Stellenbosch, South Africa. Scale: 1:1 000 000.
- Trame, Johannes & Keßler, Carsten** (Mar 2011) *Exploring the lineage of volunteered geographic information with heat maps*. In: GeoViz 2011, Hamburg, Germany.
- Treuhaft, Sarah** (2006) *The democratization of data: How the Internet is shaping the work of data intermediaries*. Working Paper 2006-03, Institute of Urban and Regional Development, University of California at Berkeley.
- Tulloch, David L** (2014) *Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice*. International Journal of Geographical Information Science, doi: 10.1080/13658816.2013.873440.
- Tyrrel, Rebecca** (27 Aug 2010) *Bruce Chatwin: Lines from a lost world*. The Telegraph.
- Uhlig, Paul F** (7 Oct 2010) *Information gulags, intellectual straightjackets, and memory holes: Three principles to guide the preservation of scientific data*. Data Science Journal, **9**:pp ES1–ES5.
- Underhill, Les G & Brooks, Michael** (2014) *Preliminary summary of changes in bird distributions between the first and second Southern African Bird Atlas Projects (SABAP1 and SABAP2)*. Ornithological Observations, **5**:pp 258–293.
- Underhill, Les G, Harebottle, Doug M & Brooks, Michael** (2012) *Second Southern African Bird Atlas Project (SABAP2): Progress report to 6 December 2012*. Ornithological Observations, **3**:pp 243–250.
- Underhill, Les G, Spiby, Jacky & Fox, Gwen** (2014) *SABAP2 shows that the Common Myna Acridotheres tristis is using the towns and villages as stepping stones to spread across South Africa*. Ornithological Observations, **5**:pp 453–456. URL <http://oo.adu.org.za/content.php?id=160>.

- United Nations Conference on Environment and Development** (3–14 Jun 1992) *Earth Summit Agenda 21: The United Nations programme of action from Rio*. Tech rep, United Nations Conference on Environment and Development, Rio de Janeiro, Brazil. URL <http://www.un.org/esa/sustdev/agenda21.htm>.
- United Nations Economic Commission for Africa** (20 Apr 2005) *Durban statement on mapping Africa for Africa*. Tech Rep E/ECA/CODI/4/36, United Nations Economic Commission for Africa. Prepared for the 4th meeting of the Committee on Development Information, Addis Ababa, Ethiopia, 23–28 April 2005, URL <http://repository.uneca.org/handle/10855/3303>.
- United Nations Economic Commission for Africa** (2009) *Scientific development, innovation and the knowledge economy*. Concept note E/ECA/CODIST/1/INF/4, United Nations Economic Commission for Africa. Prepared for the First Session of the Committee on Development Information, Science and Technology (CODIST-I), Addis Ababa, Ethiopia, 28 April – 1 May 2009, URL <http://www.uneca.org/codist/codist1/content/E-ECA-CODIST-1-INF-4-EN.pdf>.
- United Nations Economic Commission for Europe** (Feb 2009) *Common Metadata Framework, Part A: Statistical metadata in a corporate context: A guide for managers*. ECE/CES/1, United Nations Economic Commission for Europe (UNECE), Geneva, Switzerland. URL <http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework>.
- United States of America** (15 Dec 1791) *Amendments to the Constitution of the United States of America*. US Government Printing Office.
- Universität Düsseldorf** (6 Feb 2013) *Die Erklärung der Universität Düsseldorf zu Annette Schavan*. Zeit Online Hochschule. URL <http://www.zeit.de/studium/hochschule/2013-02/uni-duesseldorf-schavan-erklaerung>.
- Upadhyaya, Shyam & Todorov, Valentin** (2009) *UNIDO data quality: A quality assurance framework for UNIDO statistical activities*. Working Paper 06/2008, United Nations Industrial Development Organization (UNIDO), Vienna, Austria.
- US Court of Appeals for the Ninth Circuit** (2007) *Perfect 10, Inc v Amazon.com, Inc and A9.com Inc and Google Inc*. Case F.3d, US Court of Appeals for the Ninth Circuit. URL <http://www.ca9.uscourts.gov/datastore/opinions/2007/12/03/0655405.pdf>.
- US Court of Appeals for the Second Circuit** (7 Mar 1991) *UNITED STATES of America vs Robert Tappan MORRIS*. Case F.2d, US Court of Appeals for the Second Circuit, United States of America. URL [http://scholar.google.com/scholar\\_case?case=551386241451639668](http://scholar.google.com/scholar_case?case=551386241451639668).
- Usborne, Simon** (9 Jan 2016) *Meet the stars of Instagram: Complete unknowns who are actually very famous*. The Independent. URL <http://www.independent.co.uk/news/people/meet-the-stars-of-instagram-complete-unknowns-who-are-actually-very-famous-a6803001>.
- Vahed, Anwar** (5–9 Sep 2005) *Towards a 21st Century technology platform for improving data*

## Bibliography

---

- quality in Statistics South Africa*. In: Commonwealth Statisticians [2005], p 6. URL <http://www.statssa.gov.za/commonwealth/speakerpresentations.asp>.
- van Biljon**, Willem (29 Mar– 3 Apr 1987) *A geographical database system*. In: AutoCarto [1987], pp 356–362. URL <http://www.mapcontext.com/autocarto/proceedings/auto-carto-8/>.
- van de Groenendaal**, Hans (2014) *Gough Island emergency — amateur radio to the rescue*. EngineerIT. URL [www.ee.co.za/article/gough-island-emergency-amateur-radio-rescue](http://www.ee.co.za/article/gough-island-emergency-amateur-radio-rescue).
- Van der Walt**, Merrill V M, **Cooper**, Antony K, **Netterberg**, Inge S J & **Rubidge**, Bruce S (2015) *Putting fossils on the map: Applying a geographical information system to heritage resources*. South African Journal of Science, **111**(11/12), doi: 10.17159/sajs.2015/20140371.
- Van der Walt**, Merrill V M, **Day**, Michael, **Rubidge**, Bruce S, **Cooper**, Antony K & **Netterberg**, Inge S J (Dec 2010) *A new GIS-based biozone map of the Beaufort Group (Karoo Supergroup), South Africa*. Palaeontologia africana, **45**:pp 1–5.
- van Exel**, M, **Dias**, E & **Fruijt**, S (Feb 2011) *Proposing a redefinition of the social geographic information domain — why perpetuating the use of ‘VGI’ will lead to misconceptions and information clutter*. In: Çöltekin & Clarke [2011a], pp 29–36.
- van Genderen**, Cynthia Warringa (24 Jun 2013) *The Right to Know: A Comparative Legal Survey of Access to Official Information in Different Countries*. Master’s thesis, University of Leiden, The Netherlands.
- van Loenen**, B, **Zevenbergen**, J & **Besemer**, J (eds) (15–19 Jun 2009b) *GSDI 11 World Conference: Spatial Data Infrastructure Convergence: Building SDI Bridges to Address Global Challenges*, Rotterdam, The Netherlands.
- Van Niekerk**, Andries F & **Combrinck**, Ludwig (2012) *The use of civilian-type GPS receivers by the military and their vulnerability to jamming*. South African Journal of Science, **108**(5/6):p 4. URL <http://dx.doi.org/10.4102/sajs.v108i5/6.749>.
- Van Noorden**, Richard (17 Jan 2013) *Mathematicians aim to take publishers out of publishing*. Nature, doi: 10.1038/nature.2013.12243. URL <http://www.nature.com/news/mathematicians-aim-to-take-publishers-out-of-publishing-1.12243>.
- Van Noorden**, Richard (24 Feb 2014) *Publishers withdraw more than 120 gibberish papers*. Nature, doi: 10.1038/nature.2014.14763. URL <http://www.nature.com/news/publishers-withdraw-more-than-120-gibberish-papers-1.14763>.
- van Onselen**, Gareth (1 Jul 2013) *Sipho Seepe: A ‘sloppy’ academic unworthy to be a vice-chancellor*. Business Day. URL <http://www.bdlive.co.za/opinion/columnists/2013/07/01/sipho-seepe-a-sloppy-academic-unworthy-to-be-a-vice-chancellor>.
- van Oort**, Pepijn A J (2006) *Spatial data quality: From description to application*. PhD thesis, Wageningen University, The Netherlands.

## Bibliography

- van Zyl, Gareth** (6 Jan 2015a) *Over 30 Uber cars impounded in Cape Town*. Fin24com. URL <http://www.fin24.com/Tech/News/Over-30-Uber-cars-impounded-in-Cape-Town-20150106>.
- van Zyl, Gareth** (10 Nov 2015b) *Uber, WesBank in R200m car rental deal*. Fin24com. URL <http://www.fin24.com/Tech/News/uber-wesbank-in-r200m-car-rental-deal-20151110>.
- van Zyl, Gareth** (13 Jan 2016) *WhatsApp faces possible regulation in SA*. Fin24com. URL <http://www.fin24.com/Tech/News/whatsapp-faces-possible-regulation-in-South-Africa-20161113>.
- Vander Wal, Thomas** (2 Feb 2007) *Folksonomy coinage and definition*. Off the Top. QAccessed 4 July 2014, URL <http://vanderwal.net/folksonomy.html>.
- Vaux, David** (19 Jun 2013) *Why I retracted my Nature paper: A guest post from David Vaux about correcting the scientific record*. Retraction Watch: Tracking retractions as a window into the scientific process. URL <http://retractionwatch.com/2013/06/19/why-i-retracted-my-nature-paper-a-guest-post-from-david-vaux-about-correcting-the-scientific-record/>.
- Vecchiatto, Paul** (23 Aug 2007) *Seacom 'willing to collaborate'*. ITWeb. URL [http://www.itweb.co.za/index.php?option=com\\_content&view=article&id=7488](http://www.itweb.co.za/index.php?option=com_content&view=article&id=7488).
- Vincent, David** (Oct 2013) *Surveillance, privacy and history*. History and Policy. URL <http://www.historyandpolicy.org/papers/policy-paper-151.htm>.
- Vines, Timothy H, Albert, Arianne Y K, Andrew, Rose L, ébarre, Florence, Bock, Dan G, Franklin, Michelle T, Gilbert, Kimberly J, Moore, Jean-Sébastien & Renault, Sébastien** (19 Dec 2013) *The availability of research data declines rapidly with article age*. Current Biology, doi: 10.1016/j.cub.2013.11.014.
- Virgil (Publius Vergilius Maro)** (29 BCE) *Georgics*, book 3. Oxford University Press. Verse 284. Translated in *The Concise Oxford Dictionary of Quotations*, 2nd Ed, 1981.
- Von Drehle, David** (1 Aug 2013) *The surveillance society: Secrets are so 20th century now that we have the ability to collect and store billions of pieces of data forever*. TIME.
- Wakefield, Jane** (28 Mar 2014) *Facebook drones to offer low-cost net access*. BBC News: Technology. URL <http://www.bbc.com/news/technology-26784438>.
- Warf, Barney & Sui, Daniel** (2010) *From GIS to neogeography: ontological implications and theories of truth*. Annals of GIS, **16**(4):pp 197–209.
- Warren, Samuel D & Brandeis, Louis D** (15 Dec 1890) *The right to privacy*. Harvard Law Review, **4**(5).
- Waterford, Jack** (5 May 2013) *The library of discarded books*. Canberra Times. URL <http://www.canberratimes.com.au/comment/the-library-of-discarded-books-20130504-2izs2.html>.
- Watson, Bruce W, Venter, Fritz K & Kourie, Derrick G** (11–14 Jun 2012) *Pattern matching*

## Bibliography

---

- in structured multi-sensor/layered image big-data*. In: 33rd Canadian Symposium on Remote Sensing, Ottawa, Canada.
- Weaver**, Matthew (21 Mar 2014) *How Brown Moses exposed Syrian arms trafficking from his front room*. The Guardian. URL <http://www.theguardian.com/world/2013/mar/21/frontroom-blogger-analyses-weapons-syria-frontline>.
- Weber-Wulff**, Debora (25 Jul 2012) *Viewpoint: The spectre of plagiarism haunting europe*. BBC News. URL <http://www.bbc.co.uk/news/18962349>.
- Weiner**, Ben (25 Apr 1997) *A parody paper in solid state physics, published in 1931*. URL <http://www.math.tohoku.ac.jp/~kuroki/Sokal/misc/bethespoof.html>.
- Wesolowski**, Amy, **Buckee**, Caroline O, **Bengtsson**, Linus, **Wetter**, Erik, **Lu**, Xin & **Tatem**, Andrew J (29 Sep 2014) *Commentary: Containing the Ebola outbreak — the potential and challenge of mobile network data*. PLOS Current Outbreaks, doi: 10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e.
- Wheeler**, Tom (4 Sep 2014) *The facts and future of broadband competition*. Prepared Remarks of FCC Chairman Tom Wheeler, URL [http://transition.fcc.gov/Daily\\_Releases/Daily\\_Business/2014/db0904/DOC-329161A1.pdf](http://transition.fcc.gov/Daily_Releases/Daily_Business/2014/db0904/DOC-329161A1.pdf).
- Wiggins**, Andrea & **Crowston**, Kevin (4–7 Jan 2011) *From Conservation to Crowdsourcing: A Typology of Citizen Science*. In: *Proceedings of the 44th Hawaii International Conference on System Sciences (HICSS)*, IEEE Computer Society, pp 1–10.
- Wikipedia** (24 Jul 2010) *Reliability of Wikipedia*. Wikipedia. URL [http://en.wikipedia.org/wiki/Reliability\\_of\\_Wikipedia](http://en.wikipedia.org/wiki/Reliability_of_Wikipedia).
- Wikipedia** (18 Jan 2012) *English wikipedia blackout: From wikipedia, the free encyclopedia*. Wikipedia. URL [http://en.wikipedia.org/wiki/English\\_Wikipedia\\_blackout](http://en.wikipedia.org/wiki/English_Wikipedia_blackout).
- Wille**, Rudolf (1982) *Restructuring lattice theory: An approach based on hierarchies of concepts*. In: **Rival**, I (ed), *Ordered sets*, D Reidel Publishing Company, Dordrecht-Boston, pp 445–470.
- Wilson**, J W, **Symes**, Craig T, **Brown**, M, T, **Bonnevie Bo**, **De Swardt**, Dawid H & **Hammer**, D (2009) *A re-evaluation of morphological differences in the karoo thrush turdus smithi - olive thrush turdus olivaceus species complex*. Ostrich: Journal of African Ornithology, 80(3):pp 171–175.
- Winterman**, Denise (4 Dec 2013) *Cyber self-harm: Why do people troll themselves online?* BBC News Magazine. URL <http://www.bbc.co.uk/news/magazine-25120783>.
- WIPO (28 Sep 1979) *Berne Convention for the Protection of Literary and Artistic Works*. World Intellectual Property Organization.
- Wise**, Lindsay & **Landay**, Jonathan S (20 Jun 2013) *Government could use metadata to map your every move*. McClatchy Washington Bureau. URL <http://www.mcclatchydc.com/2013/06/20/194505/government-could-use-metadata.html>.
- Woldai**, Tsehaie (14–18 Oct 2002) *Geospatial data infrastructure: the problem of developing metadata for geoinformation in Africa*. In: 4th AARSE Conference on geoinformation for

- sustainable development in Africa*, Abuja, Nigeria. URL <http://www.itc.nl/library/papers/woldai.pdf>.
- Woloshin, Steven & Schwartz, Lisa M** (5 Jun 2006) *Media reporting on research presented at scientific meetings: more caution needed*. Medical Journal of Australia, **184**(11):pp 576–580.
- Woollaston, Victoria** (27 Mar 2014) *The app that lets you chat WITHOUT an internet or phone connection: FireChat uses hidden iOS feature to relay messages*. Mail Online. URL <http://www.dailymail.co.uk/sciencetech/article-2590589/The-app-lets-chat-WITHOUT-internet-phone-connection-FireChat-uses-hidden-iOS-feature-relay-messages.html>.
- World Bank Group** (2016) *World Development Report 2016: Digital Dividends. Overview booklet*. 102724, International Bank for Reconstruction and Development/The World Bank, Washington DC, USA, doi: 10.1596/978-1-4648-0671-1. A PDF.
- WorldWideWorx** (Dec 2012) *Executive summary: Internet access in South Africa 2012*. Tech rep, WorldWideWorx, Johannesburg, South Africa.
- Wright, Dale** (Feb 2011) *Evaluating a citizen science research programme: Understanding the people who make it possible*. Master's thesis, University of Cape Town. URL [http://adu.org.za/pdf/Wright\\_D\\_2011\\_MSc\\_thesis.pdf](http://adu.org.za/pdf/Wright_D_2011_MSc_thesis.pdf).
- Wright, Malcolm & Armstrong, J Scott** (Mar/Apr 2008) *The ombudsman: Verification of citations: Faulty towers of knowledge?* Interfaces, **38**(2):pp 125–139.
- Wu, Changxu, Zhao, Guozhen, Lin, Bin & Lee, Jonghoon** (2011) *Navigating a car in an unfamiliar country using an internet map: effects of street language formats, map orientation consistency, and gender on driver performance, workload and multitasking strategy*. Behaviour & Information Technology, doi: 10.1080/0144929X.2011.566941. Accepted.
- Wunsch-Vincent, Sacha & Vickery, Graham** (12 Apr 2007) *Participative web: User-created content*. Tech Rep DSTI/ICCP/IE(2006)7/FINAL, Organisation for Economic Co-operation and Development. Compiled for the Working Party on the Information Economy (WPIE) of the Committee for Information, Computer and Communications Policy of the OECD's Directorate for Science, Technology and Industry.
- Wyoming, Sixty-Third Legislature of the State of** (5 Mar 2015) 6-3-414. *Trespassing to unlawfully collect resource data; unlawful collection of resource data*. SF0012, The Legislature of the State of Wyoming. URL <http://legisweb.state.wy.us/2015/Enroll/SF0012.pdf>.
- Yang, Jeff** (4 Oct 2014) *Hong Kong protesters in cyberwar*. CNN. URL <http://edition.cnn.com/2014/10/03/opinion/yang-hong-kong/index.html?eref=edition>.
- Yeung, Albert KW & Hall, G Brent** (2007) *Spatial Database Systems: Design, Implementation and Project Management*. Springer, ISBN 978-1-4020-5392-4.
- Yevtushenko, Serhiy, Kaiser, Tim & Tane, Julian** (28 Jul 2003) *Concept Explorer The User Guide*.

## Bibliography

---

- Yevtushenko, Serhiy A** (2000) *System of data analysis "Concept Explorer"*. In: *Proceedings of the 7th national conference on Artificial Intelligence KII-2000*, Russia, pp 127–134. In Russian.
- Yglesias, Matthew** (13 Jan 2013) *The brilliant life and tragic death of Aaron Swartz*. Slate.com. URL [http://www.slate.com/blogs/moneybox/2013/01/13/brilliant\\_life\\_and\\_tragic\\_death\\_of\\_aaron\\_swartz.html](http://www.slate.com/blogs/moneybox/2013/01/13/brilliant_life_and_tragic_death_of_aaron_swartz.html).
- Young, Jason C & Gilmore, Michael P** (2013) *The spatial politics of affect and emotion in participatory GIS*. *Annals of the Association of American Geographers*, 103(4):pp 808–823.
- Yu, Le & Gong, Peng** (20 Jun 2012) *Google earth as a virtual globe tool for earth science applications at the global scale: progress and perspectives*. *International Journal of Remote Sensing*, 33(12):pp 3966–3986.
- Zakon, Robert H'obbes'** (1 Nov 2006) *Hobbes' Internet Timeline — the definitive ARPAnet & Internet history*. Zakon Group LLC, 8(2). URL <http://www.zakon.org/robert/internet/timeline/>.
- Zargar, Amin & Devillers, Rodolphe** (2009) *An operation-based communication of spatial data quality*. In: *2009 International Conference on Advanced Geographic Information Systems & Web Services*, IEEE, pp 140–145.
- Zennaro, M, Canessa, E, Sreenivasan, K R, Rehmatullah, A A & Cottrell, R L** (Fall 2006) *Scientific measure of Africa's connectivity*. *Information Technologies and International Development*, 3(1):pp 55–64.
- Zetter, Kim** (18 Sep 2014) *School dropout codes chat program that foils nsa spying*. Wired UK. URL <http://www.wired.co.uk/news/archive/2014-09/18/encrypted-chat>.
- Zhang, J-X & Goodchild, Michael F** (eds) (2008) *Spatial Uncertainty, Proceedings of the Eighth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, vol 1. World Academic Union, Liverpool.
- Zhou, Michelle X, Wang, Fei, Zimmerman, Thomas, Yang, Huahai, Haber, Eben & Gou, Liang** (15–19 Jul 2013) *Computational discovery of personal traits from social multimedia*. In: *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, San Jose, CA, pp 1–6, doi: 10.1109/ICMEW.2013.6618398.
- Zibi, Songezo** (12 May 2015) *Newspapers cannot provide writers with list of neutral words they may use*. Business Day. URL <http://www.bdlive.co.za/opinion/columnists/2015/05/12/newspapers-cannot-provide-writers-with-list-of-neutral-words-they-may-use>.
- Zielinski, Caroline** (25 Aug 2014) *Oh, brother: Tributes flow for David Attenborough, as Richard Attenborough dies*. The Sydney Morning Herald. URL <http://www.smh.com.au/action/printArticle?id=60839105>.
- Zielstra, D & Zipf, A** (2010) *A comparative study of proprietary geodata and volunteered geographic information for Germany*. In: *13th AGILE International Conference on Geographic Information Science*, Guimaraes, Portugal.

---

*Bibliography*

- Zielstra, Dennis, Hochmair, Hartwig H, Neis, Pascal & Tonini, Francesco** (2014) *Areal delineation of home regions from contribution and editing patterns in OpenStreetMap*. ISPRS International Journal of Geo-Information, **3**:pp 1211–1233.
- Zittrain, Jonathan** (1 Jun 2014a) *Facebook could decide an election without anyone ever finding out: The scary future of digital gerrymandering — and how to prevent it*. New Republic. URL <https://newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering>.
- Zittrain, Jonathan** (13 May 2014b) *Is the EU compelling Google to become about.me? The Future of the Internet*. URL <http://blogs.law.harvard.edu/futureoftheinternet/2014/05/13/is-the-eu-compelling-google-to-become-about-me/>.
- Zook, Matthew, Graham, Mark, Shelton, Taylor & Gorman, Sean** (2012) *Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake*. World Medical & Health Policy, **2**(2):pp 7–33.
- Zuiderwijk, Anneke, Jeffery, Keith & Janssen, Marijn** (2012) *The potential of metadata for linked open data and its value for users and publishers*. Journal of Democracy (JeDEM), **4**(2):pp 222–244.

## Web pages

*Please note that this thesis has been written from 2009 to 2016 and includes many references with URLs. Unfortunately, some of these links are now broken, and it has not been possible to verify them all again.*

Amazon (2016) *Amazon Mechanical Turk: Artificial Artificial Intelligence. Home page.* URL <https://www.mturk.com/>.

Andries Naude (2009) *andries naude. Home page.* URL <http://www.andriesnaude.co.za/>.

Animal Demography Unit (2016a) *Birds in Reserves Project: A citizen science project of the Animal Demography Unit. Home page.* URL <http://birp.adu.org.za/>.

Animal Demography Unit (2016b) *Southern African Bird Atlas Project 2. Home page.* URL <http://sabap2.adu.org.za/>.

Animal Demography Unit (2016c) *Southern African Butterfly Conservation Assessment. Home page.* URL <http://sabca.adu.org.za/>.

Animal Demography Unit (2016d) *Southern African Reptile Conservation Assessment. Home page.* URL <http://sarca.adu.org.za/>.

BBC (1986) *Domesday reloaded. home page.* URL <http://www.bbc.co.uk/history/domesday>.

BBC (2016) *BBC. Home page.* URL <http://www.bbc.co.uk/>.

BGS (2011) *British Geological Survey: Applied geoscience for our changing Earth. This site's aim is to create a geological community where geologists (whether amateur or professional) can report a geological observation.* URL <https://britishgeologicalsurvey.crowdmap.com/main>.

Bing (2016) *Bing Maps. Home page.* URL <http://maps.bing.com/maps/>.

Butterfly Conservation (2016) *Big Butterfly Count. Home page.* URL <http://www.bigbutterflycount.org/>.

- CGIS (1 Jun 2011) *SDI workshop held in Cape Town*. URL <http://web.up.ac.za/default.asp?ipkCategoryID=16055&subid=16055&ipklookid=11&archive=1&ArticleID=7512>.
- CODATA Task Group on Data Citation Standards and Practices (2014) *Codata task group on data citation standards and practices: Terms of reference*. URL <http://www.codata.org/taskgroups/TGdatacitation/index.html>.
- Compete (2016) *Compete: Know where you stand. Home page*. URL <http://www.compete.com/>.
- Cornell Lab of Ornithology (2016) *ebird: An online database of bird distribution and abundance. home page*. URL <http://ebird.org/>.
- Creative Commons (2016) *Creative commons. home page*. URL <http://creativecommons.org/>.
- Cryptome (2016) *Cryptome. home page*. URL <http://www.cryptome.org/>.
- CyberTracker (2016) *CyberTracker. Home page*. URL <http://www.cybertracker.co.za/>.
- D-Lib (2016) *D-Lib Magazine: The magazine of digital library research. Home page*. doi: 10.1045/dlib.magazine. URL <http://www.dlib.org/>.
- DCMI (2016) *Dublin Core Metadata Initiative (DCMI). Home page*. URL <http://dublincore.org/>.
- Delicious (2016) *Delicious. Home page*. Formerly known as *del.icio.us.*, URL <http://delicious.com/>.
- DOI (2016) *The DOI System. Home page*. URL <http://doi.org/>.
- Eskom (2016) *Eskom Expo for Young Scientists. Home page*. URL <http://www.exposcience.co.za/>.
- European Union (2016) *EUOSME. Home page*. URL <http://www.inspire-geoportal.eu/EUOSME/>.
- Facebook (2016) *Facebook. Home page*. URL <http://www.facebook.com/>.
- FGDC metadata (2016) *Metadata. Need to update this reference, particularly the page title*, URL <http://www.fgdc.gov/metadata/>.
- FidoNet (2016) *FidoNet. Home page*. URL <http://www.fidonet.org/>.
- First Monday (2016) *First Monday: Peer-reviewed journal on the Internet. Home page*. URL <http://firstmonday.org/>.
- Flickr (2016) *Flickr. Share your photos. Watch the world. Home page*. URL <http://www.flickr.com/>.
- Foldit (2016) *Foldit: Solve puzzles for science. Home page*. URL <http://fold.it/portal/>.
- Free Software Foundation (2016) *Gnu general public license*. URL <https://www.gnu.org/licenses/gpl.html>.

## Web pages

---

- FrontlineSMS (2016) *FrontlineSMS. Home page*. URL <http://www.frontlinesms.com/>.
- Geeknet, Inc (2016) *SourceForge: Find, Create, and Publish Open Source software for free. Home page*. URL <http://sourceforge.net/>.
- Genealogical Society of South Africa (2016) *The Genealogical Society of South Africa: Our Business is Genealogy . Home page*. URL <http://www.genza.org.za/>.
- Geonetwork (2016) *Geonetwork opensource. home page*. URL <http://geonetwork-opensource.org/>.
- GitHub, Inc (2016) *GitHub: How people build software. Home page*. URL <https://github.com/>.
- Google (2013) *Loon for All — Project Loon — Google. Home page*. Accessed 17 June 2013 at; URL <http://www.google.com/loon/>.
- Google (2016a) *Google Earth: Explore, Search, and Discover. Home page*. URL <http://earth.google.com/>.
- Google (2016b) *Google. Home page*. URL <http://www.google.com/>.
- Google (2016c) *Google Map Maker. Home page*. URL <http://www.google.com/mapmaker>.
- Google (2016d) *Google maps. Home page*. URL <http://maps.google.com/>.
- Google (2016e) *Google Scholar Beta: Stand on the shoulders of giants. Home page*. URL <http://scholar.google.com/>.
- Google (2016f) *Google Street View. Home page*. URL <http://www.google.com/maps/streetview/>.
- Google (2016g) *Google Trends. Home page*. URL <http://www.google.com/trends>.
- Google (2016h) *Google.org Flu Trends. Home page*. URL <http://www.google.org/flutrends/>.
- Gormley, Antony (6 Jul – 14 Oct 2009) *One & Another. Home page*. URL <http://www.oneandother.co.uk/>.
- Goworldwind (2016) *goworldwind.org: Complete Guide to World Wind Help Resources. Home page*. URL <http://goworldwind.org/>.
- Gwent Police (2016) *Heddlu Gwent Police: COW — The film that will stop you txting and driving. Web page*. URL <http://www.gwent.police.uk/leadnews.php?a=2172>.
- HarassMap (2016) *HarassMap. Home page*. URL <http://harassmap.org/>.
- HERE Map Creator (2016) *Map your world near and far with HERE Map Creator. Home page*. URL <http://mapcreator.here.com/mapcreator>.
- HOT (2016) *Humanitarian OpenStreetMap Team. Home page*. URL <http://hot.openstreetmap.org/>.
- Hudma (2016) *Huduma — Fix my constituency! Home page*. URL <http://huduma.info/>.

- IANA (2016) *Iana character set register*. URL <http://www.iana.org/assignments/character-sets>.
- IEEE LTSC (2016) *Ieee wg12: Learning object metadata — ltsc home. web page*. URL <http://ltsc.ieee.org/wg12/>.
- IMF (2016) *Data Quality Reference Site. Web page*. URL <http://dsbb.imf.org/Pages/DQRS/DQAF.aspx>.
- InnoCentive (2016) *Welcome to InnoCentive: Where the world innovates. Web page*. URL <http://www.innocentive.com/>.
- International Cartographic Association (16 Aug 2003) *Mission*. URL <http://icaci.org/mission/>.
- International Labour Organization (2016) *Isc0: International standard classification of occupations*. URL <http://www.ilo.org/public/english/bureau/stat/isco/index.htm>.
- International Neuroethics Society (2016) *International neuroethics society: Advancing the development and responsible application of neuroscience. home page*. URL <http://ns.memberclicks.net/>.
- Internet Archive (2016a) *Internet archive: Universal access to all knowledge. home page*. URL <http://archive.org/>.
- Internet Archive (2016b) *The wayback machine. home page*. URL <http://archive.org/web/web.php>.
- ISO (2016b) *ISO Concept Database. Home page*. URL <http://cdb.iso.org/>.
- ISO/TC 211 (2016) *ISO/TC 211, Geographic information/Geomatics. Home page*. URL <http://www.isotc211.org/>.
- Kiva (2016) *Kiva: Empower people around the world with a \$25 loan. Home page*. URL <http://www.kiva.org/>.
- Leiner, Barry M and Cerf, Vinton G and Clark, David D and Kahn, Robert E and Kleinrock, Leonard and Lynch, Daniel C and Postel, Jon and Roberts, Larry G and Wolff, Stephen (10 Dec 2003) *Histories of the internet: A brief history of the internet*. URL <http://www.isoc.org/internet/history/brief.shtml>.
- LINZ (2016) *Building Our Footprints. Home page*. URL <http://canterburymaps.govt.nz/BuildingOurFootprints/>.
- LocalBlock (2016) *LocalBlock: Connect with your neighbourhood. Home page*. URL <https://localblock.co.za/>.
- Lynx Bird Ticks (2016) *Lynx BirdTicks Southern Africa: Your ultimate birding companion. Home page*. URL <http://lynxnature.com/>.
- Mobilitate (2015) *Mobilitate, a better South Africa. Home page*. No longer exists, URL <http://www.mobilitate.co.za/>.

## Web pages

---

- Myspace (2016) *Myspace: Welcome to the neighborhood. Home page.* URL <http://myspace.com/>.
- NASA (2016) *NASA World Wind: Open source \* cross platform \* open standards. Home page.* URL <http://worldwind.arc.nasa.gov/>.
- National Audubon Society (2016) *The Christmas bird count: History. Web page.* URL <http://www.audubon.org/bird/cbc/history.html>.
- National Geographic (2016) *Field expedition: Mongolia. home page.* URL <http://exploration.nationalgeographic.com/mongolia>.
- NaturalWorld (2016) *NaturalWorld. Home page.* URL <http://www.natworld.org/>.
- Navteq (2016) *NAVTEQ. Home page.* URL <http://www.navteq.com/>.
- Newdea (2016) *newdea™: Overview. web page.* URL <http://newdea.com/our-approach/overview>.
- NGI (2016) *NGI — National Geo-spatial Information (NGI). Mapping our nation. Home page.* URL <http://www.ngi.gov.za/>.
- OAIC (2016) *What is privacy? URL* <http://www.privacy.gov.au/aboutprivacy/what>.
- OGC (2016) *Open Geospatial Consortium, Inc (OGC). Home page.* URL <http://www.opengeospatial.org/>.
- Okoli, Chitu and Mehdi, Mohamad and Mesgari, Mostafa and Nielsen, Finn Årup and Lanamäki, Arto (2016) *WikiLit: A literature review of scholarly research on Wikipedia.* URL [http://wikilit.referata.com/wiki/Main\\_Page](http://wikilit.referata.com/wiki/Main_Page).
- Old Weather (2016) *Old weather: Our weather's past, the climate's future. Home page.* URL <http://www.oldweather.org/>.
- Open Definition (2016) *Open Definition. Home page.* URL <http://opendefinition.org/>.
- OpenAddresses (2016) *OpenAddresses: Management of free and open worldwide localized addresses. Home page.* URL <http://www.openaddresses.org/>.
- OpenStreetMap (2016) *OpenStreetMap: The Free Wiki World Map. Home page.* URL <http://www.openstreetmap.org/>.
- Ordnance Survey (2016) *Positional accuracy improvement programme.* URL <http://www.ordnancesurvey.co.uk/oswebsite/pai/faqgeneral.html>.
- Oxford (2016) *Oxford dictionaries: The world's most trusted dictionaries. home page.* URL <http://oxforddictionaries.com/>.
- Oxford Dictionaries (2016) *Oxford Dictionaries — The world's most trusted dictionaries.* URL <http://oxforddictionaries.com/>.
- Padkos (2016) *Padkos. Home page.* URL <http://www.padkos.co.za/>.
- Panoramio (2016) *Panoramio. Photos of the world. Home page.* URL <http://www.panoramio.com/>.

- Peattie, Charles and Taylor, Russell (2016) *Alex. home page*. URL <http://www.alexcartoon.com/>.
- Peer to Patent (2016) *Peer to Patent: Community Patent Review. Home page*. In cooperation with the United States Patent and Trademark Office, URL <http://www.peertopatent.org/>.
- Planet Hunters (2016) *Planet hunters. Home page*. URL <http://www.planethunters.org/>.
- Precinct Web (2016) *Precinct Web. Home page*. URL <http://www.precinctweb.com/>.
- QGIS (2014) *QGIS: A Free and Open Source Geographic Information System*. URL <http://www.qgis.org/>.
- Reporters Without Borders (2013) *Reporters Without Borders: Press Freedom Barometer 2013. Home page*. URL <http://en.rsrf.org/press-freedom-barometer-netizens-imprisoned.html>.
- SeeClickFix (2016) *SeeClickFix. Home page*. URL <http://www.seeclickfix.com/>.
- SETI@Home (2016) *SETI@Home. Home page*. URL <http://setiathome.berkeley.edu/>.
- South African Geographical Names Council (SAGNC) (2016) *Official South African Geographical Names System. Home page*. URL <http://sagns.dac.gov.za/>.
- Speakers' Corner Trust (2016) *Speakers' Corner Trust. Home page*. URL <http://www.speakerscornertrust.org/>.
- Spisys (2016) *Welcome to Spisys: Spatial planning and information. Home page*. URL <http://fs.spisys.gov.za/>.
- Stichting Spreeksteen Amsterdam (2016) *Spreeksteen.nl: Voor het openbaren van gedachten of gevoelens [...] heeft niemand voorafgaand verlof nodig [...] Grondwet, art 7.3*. URL <http://www.spreeksteen.nl/>.
- Tame, Joseph (2011) *Joseph tame: Tokyo based social-mobile tech creator. art of running — a tribute to steve: 21km apple logo*. URL <http://josephtha.me/2011/08/art-of-running-tokyo-steve-jobs-apple-logo/>.
- Technorati, Inc (2016) *Technorati. Home page*. URL <http://www.technorati.org/>.
- TomTom (2016) *Tom Tom: Portable GPS car navigation systems. Home page*. URL <http://www.tomtom.com/>.
- Tracks4Africa (2016) *Tracks4Africa: Mapping Africa, one day at a time. Home page*. URL <http://www.tracks4africa.co.za/>.
- Twitter (2016) *Welcome to Twitter. Start a conversation, explore your interests, and be in the know. Home page*. URL <http://www.twitter.org/>.
- United Nations Economic Commission for Europe (2016) *Statistical metadata (METIS). Home page*. URL <http://live.unece.org/stats/archive/04.01d.e.html>.

## Web pages

---

- United Nations Statistics Division (2016) *Committee for the Coordination of Statistical Activities. Home page*. URL [http://unstats.un.org/unsd/acsub-public/workpartner\\_ccsa.htm](http://unstats.un.org/unsd/acsub-public/workpartner_ccsa.htm).
- Ushahidi (2016) *Ushahidi: Changing the world one map at a time. home page*. URL <http://www.ushahidi.com/>.
- W3C (2016) *Semantic Web. Home page*. URL <http://www.w3.org/standards/semanticweb/>.
- Water Shortage South Africa (2016) *Water Shortage SA: Everyone can be a helper, a hero — it's easy. Home page*. URL <https://sites.google.com/site/watershortagesouthafrica/>.
- WEF (2016) *World water monitoring day. home page*. URL <http://www.worldwatermonitoringday.org/index.html>.
- WESSA (27 Nov 2011) *Friends Groups. Web page*. URL <http://wessa.org.za/get-involved/friends-groups-2.htm>.
- Wikileaks (2016) *Wikileaks. Home page*. URL <https://secure.wikileaks.org/>.
- Wikimapia (2016) *Wikimapia: Let's describe the whole world! Home page*. URL <http://wikimapia.org/>.
- Wikimedia (2016) *Wikipedia. Home page*. URL <http://en.wikipedia.org/>.
- WolframAlpha (2016) *WolframAlpha: computational knowledge engine. Home page*. URL <http://www.wolframalpha.com/>.
- X PRIZE Foundation (2016) *X PRIZE Foundation. Revolution Through Competition*. URL <http://www.xprize.org/>.
- XBRL International (2016) *XBRL: eXtensible Business Reporting Language. Home page*. URL <http://www.xbrl.org/>.
- Yahoo! (2016) *Yahoo! Maps, Driving Directions and Traffic. Home page*. URL <http://maps.yahoo.com/>.
- YouTube (2016) *YouTube: Broadcast Yourself. Home page*. URL <http://www.youtube.com/>.



---

# Colophon

A colophon is used to provide details of the production of a document, which are given below. However, a colophon is also used to explain the style of the document, which I do first. While my use of English might appear idiosyncratic to some, I think that my grasp of the language is sufficient to have my own style. Specifically, please note the following.

- For some decades now, *full stops* have not been used for many types of *abbreviations* and *acronyms* (for example, see Fowler & Gowers [1965]; Oxford English Dictionary Department [1981]). Excessive punctuation looks messy and, I believe, impedes reading. Modern usage has also tended to reducing punctuation, so I do not use any full stops in abbreviations and acronyms.
- I also do not use full stops in *eg* (*exempli gratia*) and *ie* (*id est*): as they precede lists or clarifications, or as Fowler & Gowers [1965] put it, “*that of delivering the goods that have been invoiced in the preceding words*”, I use a colon after them. However, in running text it is better to use *for example* and *that is* [Oxford English Dictionary Department 1981], which I have done.
- Similarly, I do not use a full stop immediately before a closing parenthesis and I have tried to avoid having parenthetical full sentences.
- A modern trend is to capitalize only the initial letter of an acronym when the acronym can be pronounced, giving the likes of *HIV/Aids*. This is silly, because it assumes universal pronunciation. For example, some Afrikaners pronounce *GIS* as *gis* (as in the Afrikaans word for yeast), while English-speaking South Africans tend to spell out the letters. The acronym for the Food and Agriculture Organization of the United Nations is generally spelled out, as *FAO* or *UN FAO*, but some preferring to treat the organization facetiously call it *fao*. Obviously, there are words such as *radar* and *laser* which are rarely treated as acronyms. I capitalize acronyms, to be consistent and to reduce ambiguity.
- I neither accept nor reject the Oxford comma slavishly. Generally, I do not use it, but it is convenient in a long list and/or where there might be ambiguity between items in the list. I use a semicolon where the commas in a list item might cause confusion.
- Some feel that a thesis should consist of chunks of narrative, rather than bulleted or enumerated lists. However, I have found it difficult for myself to find things in the text of this thesis when they are hidden in the narrative. Hence, bulleted

and enumerated lists are definitively and unambiguously better than chunks of narrative. Generally, I use bullets unless the order of the items is significant. In any case, as this thesis is about classification, it is appropriate for me to be classifying text through bulleted and enumerated lists!

- I find it difficult to decide how often a citation should be given in a paragraph, particularly when an exact quote is taken from the reference early in the paragraph or the author(s) contributed initially to only the first sentence of the paragraph. Repeating the citation at the end shows that the whole paragraph drew on that author. On the other hand, if some other author gets cited halfway through the paragraph, then the influence of the first author implicitly ends with the second citation, as the second author takes over. So, I prefer the more conservative approach, of repeating the citation at the end of the paragraph to ensure the cited author gets sufficient credit.

As the University of Pretoria does not have an official style for PhD theses, each student has freedom over the style they use for their text. Being sceptical about the utility of word processor packages for producing satisfactorily a large and complex document such as a thesis, I chose to use  $\text{\TeX}$  [Knuth 1984] and  $\text{\LaTeX}$  [Lamport 1986], as I did for my MSc dissertation [Cooper 1993], also at the University of Pretoria. For this thesis, I chose to use  $\text{\BIBTeX}$  [Patashnik 1988], a pre-processor that generates `\bibitem` entries from a  $\text{\BIBTeX}$  input file.  $\text{\BIBTeX}$  allows one to make changes to the formatting of the bibliographic entries *en masse* and to treat all one's references as a reference database, drawing into a document only those references that are actually used. Google Scholar [Google 2016e] makes available  $\text{\BIBTeX}$  data for all the documents it lists, though unfortunately in my experience, many of these  $\text{\BIBTeX}$  entries are incorrect — but it is the thought that counts!

$\text{\BIBTeX}$  has a limited set of *entry types*: besides the obvious omissions of electronic documents, it does cater at all for legislation, case law, patents, standards, maps or data sets. So, I used Natbib [Daly 2006] to fix up some of the limitations (eg: allowing author-date citations), and Custom-bib [Daly 2003] to customize  $\text{\BIBTeX}$ 's input files to some extent and fix up other limitations (eg: sequencing surnames and given names).

\*\*\*\*