Visualisation et sciences humaines et sociales

Les travaux de ce manuscrit ont été réalisés dans le cadre du projet BLIZAAR, un projet collaboratif international issu d'un partenariat entre la France et Luxembourg financé par l'ANR et le FNR (https://blizaar.list.lu/doku.php). Regroupant plusieurs acteurs provenant des domaines de l'informatique, des humanités numériques (voir section suivante) et de la biologie, il a pour objectif l'analyse et la compréhension des ensembles de données modélisables par des réseaux multi-couches ainsi que la création de nouvelles représentations interactives pour les réseaux multi-couches à travers la combinaison ou la création de nouvelles méthodes de visualisation. Les travaux réalisés dans ce manuscrit ont été effectués en étroite collaboration avec des experts des données à travers des discussions régulières afin de comprendre leur domaine, leurs problématiques et les solutions à mettre en oeuvre pour y répondre.

Dans ce chapitre, nous détaillons le domaine et l'activité de nos experts des données afin de pouvoir expliquer leurs objectifs (Section 2.1). Nous décrivons ensuite les données des experts et leur modélisation (Section 2.2) ainsi que les différentes similitudes rencontrées avec nos expériences passées (Section 2.3) afin de conclure sur les questions relatives aux contraintes et spécificités qui trouvent réponses dans les chapitres suivants de ce manuscrit (Section 2.4).

2.1 Contexte et objectifs

Les travaux de ce manuscrit ont été réalisés conjointement avec les historiens du C2DH (Centre for Contemporary and Digital History), oeuvrant dans le domaine de recherche des humanités numériques (Digital Humanities). La défini-

tion de ce domaine, à l'instar du "Big Data", peine à faire consensus parmi ses membres. Un historien, Fred Gibbs, écrit même en introduction d'un de ses travaux : "S'il y a deux choses dont le milieu universitaire n'a pas besoin, ce sont un autre livre sur Darwin et un autre post de blog sur la définition des humanités numériques." [32]. On peut cependant le définir d'une manière générale par "l'application du « savoir-faire des technologies de l'information [et de l'informatique/infosciences] aux questions de sciences humaines et sociales»" (wikipédia : https://fr.wikipedia.org/wiki/Humanités_numériques). L'un des rôles majeurs attribués aux humanités numériques est celui de la préservation et la valorisation du patrimoine culturel numérique : il consiste en la collection, la préservation, l'analyse et la diffusion au plus grand nombre des oeuvres numériques fabriquées par l'homme au cours du temps et en lien avec notre patrimoine culturel et historique. C'est ce que proposent les historiens avec le CVCE (Centre Virtuel de la Connaissance sur l'Europe, https://www.cvce.eu/), une infrastructure de recherche exploitant et mettant à disposition des milliers de documents de tout genre : image (photos, caricatures, affiches, etc.), vidéo (films d'archive, séminaires, interviews, etc.), son (discours, témoignages, interviews, etc.) et texte (articles, rapports, lettres, etc.) dont l'objectif est de pouvoir retracer le processus de la construction européenne. Ces données sont presque intégralement consultables sur le site du CVCE.

À partir de ce vaste corpus de documents, les historiens du CVCE composent des "ePublications", des articles en ligne sur des thématiques ou périodes historiques précises comprenant d'une part un texte descriptif ou analytique du sujet et d'autre part un ensemble de documents utilisés comme sources et/ou comme compléments d'information (https://www.cvce.eu/epublications). La conception de ces ePublications nécessite cependant de trouver les sources nécessaires. Ceci va donc être l'objectif phare des experts des données : **comment naviguer dans un vaste corpus de documents hétérogènes afin de constituer une ensemble bibliographique satisfaisant ?** Le terme satisfaisant est important car à cet objectif s'ajoutent des contraintes afin de considérer pertinent l'ensemble bibliographique : les documents utilisés doivent couvrir les éléments majeurs reconnus dans le domaine mais les documents plus marginaux (i.e. moins connus/utilisés) sont aussi valorisés et peuvent être considérés comme très intéressants. En plus de cela, une seconde contrainte est l'homogénéité de type dans l'ensemble bibliographique : la sélection des documents doit couvrir de manière équitable, dans la mesure du possible, les différents types de document existants i.e. avoir une représentation non déséquilibrée des différents types dans l'ensemble bibliographique.

Outre la conception d'ePublications, un autre objectif assez similaire est d'offrir à des utilisateurs tiers (visiteurs du site du CVCE par exemple) la possibilité de naviguer simplement et intuitivement dans le corpus afin de prendre connaissance des différents thèmes et documents majeurs qui s'y trouvent avec, encore une fois, la nécessité de pouvoir présenter des documents marginaux ou en marge si certaines pistes ou thèmes sont approfondis.

Notre but va donc être de proposer une méthode permettant de répondre simultanément aux différents objectifs des experts en offrant la possibilité de naviguer dans le corpus et de l'explorer tout en facilitant aux experts le respect des contraintes s'imposant à la conception de la base bibliographique. Dans la section suivante, nous décrivons nos données et le système complexe qu'elles forment, modélisable par un réseau multi-couche, puis les différentes expériences rencontrées relatives à ces objets dans d'autres domaines des sciences humaines et sociales.

2.2 Données et modélisation

Comme vu précédemment, l'essentiel des données utilisées par les experts correspond à un vaste corpus de documents de types variés. En plus de cela, des fiches de méta-données renseignent diverses informations dont la nature varie en fonction des types de documents (le nom des auteurs pour un écrit ou une illustration, des informations géographiques pour un lieu ou un rapport de meeting, des informations temporelles, etc.). Ces informations ne sont pas renseignées systématiquement ainsi une majorité de documents ne sont par exemple pas datés.

A partir de ces informations, une base de données a été générée référençant non seulement les ePublications et les documents du corpus mais aussi les différentes "entités" liées à ce corpus : personnes, lieux, institutions, organisations ou groupes sociaux. Pour cela, ces "entités" sont extraites des documents en utilisant essentiellement un calcul de co-occurence des mots basé sur le coefficient de Jaccard [37] au sein des documents ainsi qu'à partir des méta-données disponibles. DBpedia (https://wiki.dbpedia.org/) est aussi utilisé afin d'enrichir certaines entités lorsqu'un parallèle est possible entre l'entité et DBpedia. En plus de cela, les relations entre les différents documents et entités sont conservées : un lien est présent entre une entité et un document dans lequel elle apparaît, entre deux entités apparaissant au sein du même document, entre une ePublications et les documents sur lesquels elle s'appuie, entre deux documents partageant un même document, etc. (voir Fig. 2.1). Il en résulte un réseau composé de 51798 sommets (documents, entités et ePublications) et 1 074 643 liens.

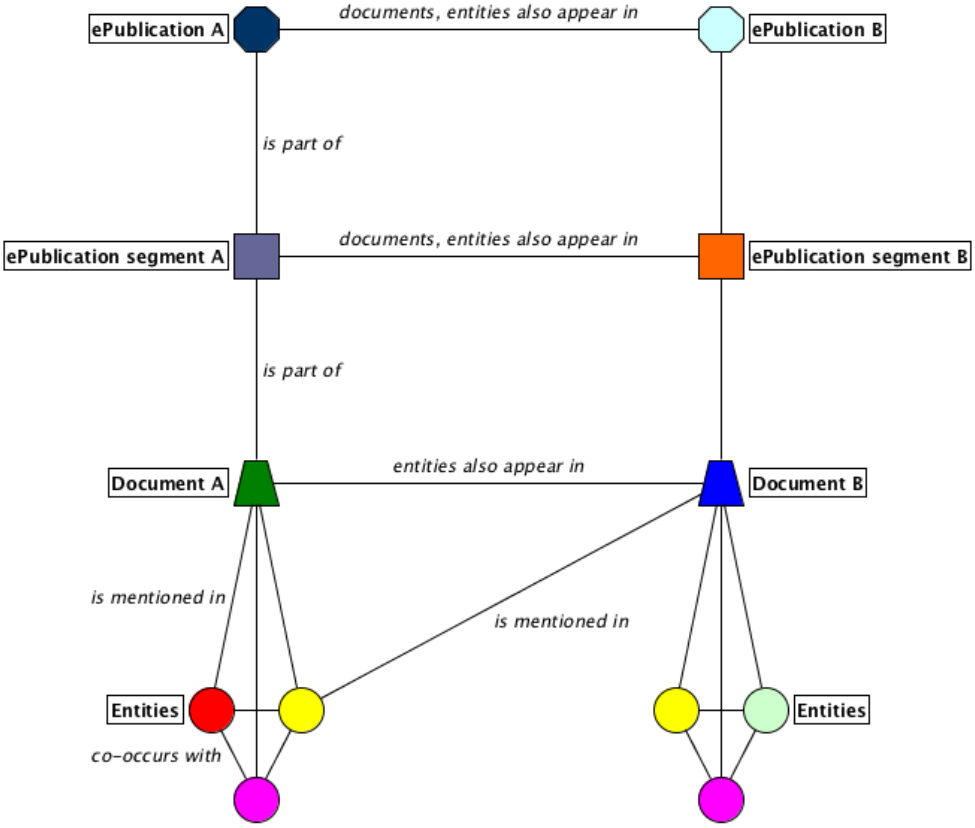

FIGURE *2.1* – ***Structure des données*** *: Les ePublications, documents et entités sont reliés entre eux. La sémantique des liens varie en fonction des types des deux éléments reliés (illustration extraite du site du projet :* `https://blizaar.list.lu/doku.php?id=BLIZAAR` *).*

Ces données forment alors un système complexe (Fig. 2.2) modélisable comme un ensemble de sous-systèmes (ou couches) inter-connectés : un réseau multi-couche. Un réseau multi-couche est défini par un ensemble de sommets (ici les documents, entités et ePublications) et un ensemble de liens (les relations entre les documents et entités) mais propose en plus un ensemble de couches i.e. des ensembles de sommets ou liens définis par les différents types possibles (l'ensemble de sommets représentant des personnes va définir la couche personne, l'ensemble des sommets représentant des localisations va définir la couche localisation, etc.) [41, 56]. Ces couches rendent complexe une analyse claire de la topologie du réseau tant la distribution peut varier en fonction des couches ou des liens inter-

FIGURE *2.2* – ***Vue générale des données*** *: Représentation noeud-lien des données de nos collègues historiens provenant de la base de données du CVCE (Centre Virtuel de Connaissance sur l'Europe). Réseau complexe de 51798 sommets et 1074643 arêtes, sa distribution est essentiellement multipolaire et hiérarchique. Un tel ensemble de données, à cause de la surcharge visuelle, de la superposition des sommets et de la complexité sémantique, impose le développement de méthodes spécialisées pour pouvoir mener une analyse efficace et pertinente. La conception et l'implémentation de ces méthodes sont développées dans les chapitres suivants.*

couches. Dans le cas présent, malgré une construction hiérarchique des données, la distribution est essentiellement multi-polaire avec des points centraux (institutions

majeures, personnalités majeures de l'Europe, etc.) inter-connectés par des sommets intermédiaires appartenant à des couches diverses. Les réseaux multi-couches sont donc de véritables mille-feuilles sémantiques nécessitant alors des méthodes propres pour être exploités pleinement.

Enfin, il est à noter que les historiens du CVCE sont déjà familiarisés avec la notion de réseaux, ce qui a permis une étroite collaboration dans la conception des méthodes présentées dans la suite. Des travaux antérieurs au projet BLIZAAR ont permis à nos collègues historiens de créer un outil, Histograph [24], permettant de filtrer et visualiser leurs données en utilisant des vues noeuds-liens. La plupart des visualisations de réseaux actuelles en humanité numérique sont publiées dans "la tradition de l'ère d'imprimerie : sous forme d'images statiques expliquées par du texte et une légende". Il a été constaté que ces visualisations peuvent être "traîtresses" car "exigent que les personnes comprennent l'objectif visé, la conceptualisation des données et les biais potentiels". Histograph a donc été développé dans l'objectif de répondre à cette observation. Permettant des visualisations proposant animations, filtres et informations contextuelles, il permet de faciliter la compréhension générale et le développement d'analyses plus efficacement qu'avec une "image statique". Dans la conclusion de ces travaux, les historiens constatent la présence d'"opportunités pour les projets d'analyse de réseau". Le projet BLIZAAR s'inscrit en continuité de cette dynamique : proposer cette fois une méthode permettant l'exploration des données, la gestion du caractère multi-couche du réseau et la prise en compte des objectifs et contraintes des experts lors de la navigation.

2.3 Travaux préliminaires

Ce n'est pas la première fois que nous exploitons des réseaux multi-couches dans le cadre des sciences humaines et sociales. Nos expériences passées nous ont amenés à travailler avec différents spécialistes (géographes, géomaticiens, juristes et sociologues) à travers deux projets, GEOBS (2015-2017) et TETRUM (2016), ayant chacun eu un impact sur la conception de la méthode `M-QuBE`^{3}. C'est notamment ces travaux qui nous ont amenés progressivement à considérer ces données comme des réseaux multi-couches et a comprendre les spécificités de ce modèle.

GEOBS

Ces dernières années, le développement des moyens de diffusion de l'information numérique, l'amélioration des techniques de géolocalisation et l'utilisation accrue

des informations environnementales par les politiques publiques ont résulté en une augmentation considérable des flux d'informations géographiques. Pour contrôler ces flux, de nombreux investissements ont été débloqués ces dernières années par les autorités publiques afin de créer des structures spécialisées : les Infrastructures de Données Géographiques (Code de l'environnement - Article L127-1) ou "IDG".

Le projet région Aquitaine GEOBS (http://geobs.cnrs.fr/) avait pour objectif d'analyser d'une part les flux d'informations transitant à travers et entre les IDG (une IDG peut partager ou moissonner les données d'autres IDG) et d'autre part les usages et moyens mis en oeuvre autour de ces plateformes. De manière simplifiée, ces IDG se présentent sous la forme d'un site internet à partir duquel il est possible d'accéder aux différentes informations et études géographiques qui y sont stockées. Ces données sont toutes accompagnées d'une fiche de méta-données renseignant la zone géographique concernée par la donnée (l'emprise), les thèmes de l'étude, les auteurs, etc. Nos travaux [62] ont consisté à utiliser ces méta-données afin de pouvoir modéliser et exploiter différents graphes permettant ainsi à nos collègues géographes d'analyser la qualité et la circulation des informations intra et inter IDG. Plusieurs approches ont donc été réalisées afin de répondre aux questions de nos experts notamment :

- Une analyse de la couverture thématique basée sur des calculs de similarité entre les mots clés et descriptifs – Est-ce que les thèmes sont équitablement répartis entre les données ? Est-ce qu'il existe des communautés thématiques majeures ? (Fig. 2.3)
- Une analyse de la gouvernance des données basée sur les différentes informations relatives aux acteurs ayant généré les données – Quels sont les acteurs phares dans le milieu ? Y a t-il des groupes d'acteurs en concurrence ou coopération ?
- Une analyse de la couverture spatiale utilisant les informations de géolocalisation – Y a t-il une homogénéité des emprises géographiques dans le territoire ? Quel est le degré de superposition des emprises des différentes études ? (Fig. 2.4)

Tous les travaux réalisés, y compris les exemples précédents, ont un point commun : chaque analyse est basée sur une métrique spécifique utilisant des attributs différents du même jeu de données. Autrement dit, chaque nouvelle analyse a nécessité d'ajuster la manière de calculer un score pour s'adapter à de nouvelles informations sémantiques issues du même jeu de données. Chaque métrique y détermine un score qui est comparé à une valeur seuil définie par l'utilisateur. Ce

seuil permet ainsi de filtrer les données afin de ne conserver que celles considérées représentatives ou intéressantes pour l'utilisateur. Par exemple, pour le graphe de similarité (Fig. 2.3), un lien n'est affiché qu'à partir d'un pourcentage de ressemblance des thèmes traités. Un seuil maximal ne va alors afficher dans le graphe que les arêtes entre des études géographiques identiques ou redondantes alors qu'un seuil nul va générer un graphe avec l'ensemble des liens (dont la sur-abondance n'est ni représentative de l'objectif ni exploitable). Pour le graphe de couverture spatiale (Fig. 2.4), le schéma est identique mais l'analyse étant centrée sur la superposition des surfaces couvertes par les études géographiques, le seuil est défini en fonction de l'intersection spatiale entre deux études. Ainsi, deux études sont liées si elles ont une surface en commun supérieure à la valeur fixée par le seuil. Le projet GEOBS a continué à posteriori du début du projet BLIZAAR et a généré plusieurs autres publications, notamment sur la communication et l'usage utilisateur des IDG [28].

Si au moment de ces travaux, nous n'avions pas encore de focus sur les graphes multi-couches, le cadre est pourtant comparable : les différents liens entre les sommets définissent des couches (similarité sémantique, superposition spatiale, gouvernance pour les trois exemples ci-dessus) à partir desquelles il est nécessaire de faire ressortir ce qui est intéressant pour l'utilisateur, comme pour les données du CVCE. De ces travaux, nous avons donc tiré deux enseignements ré-exploités lors de la conception de `M-QuBE`[3] : la nécessité de différencier le traitement pour l'adapter à chaque "couche" sémantique d'un même réseau / jeu de données (liens thématiques, liens d'appartenance, liens spatiaux...) et la nécessité de restreindre la visualisation à ce qui est le plus pertinent pour l'utilisateur, en évaluant et comparant les éléments traités en fonction des objectifs définis.

TETRUM

Les réseaux de traite des humains ne sont pas nouveaux mais ont subi un changement dû aux nouvelles techniques de communication et de partage des informations. Avec le développement d'internet, c'est toutes les pratiques et stratégies criminelles qui ont évolué. Un projet interdisciplinaire (PEPS/IdEx) comprenant juristes, sociologues et informaticiens a donc été mis en place afin d'analyser et comprendre les nouvelles formes, usages et modes opératoires de ces réseaux criminels [47] (et a été mentionné dans un article du Figaro : `https://www.labri.fr/images/uploads/Art%20Figaro-Trafic%20%EAtre%20humains`).

Contrairement aux projets décrits précédemment, les données initiales dont

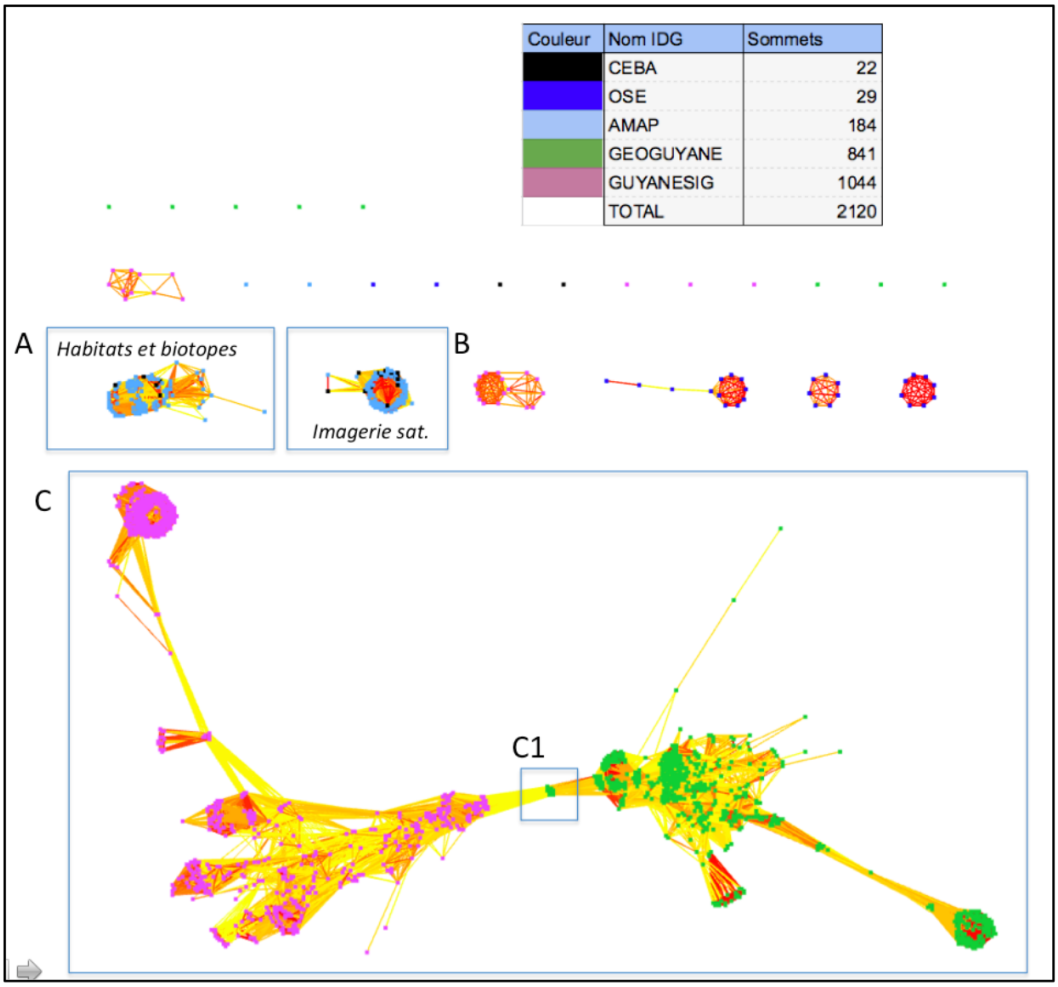

FIGURE *2.3* – ***Graphe de similarité*** *: Ce graphe est construit à partir d'un calcul de similarité entre les différentes données de cinq IDG en utilisant les mots-clés, thèmes et descriptifs contenus dans les méta-données. Les sommets représentent les études géographiques et une arête entre deux sommets indique qu'ils ont un score de similarité sémantique supérieur au seuil défini par les experts (arête jaune : similarité minimale, arête rouge : similarité maximale). Les différentes communautés représentent alors des groupements thématiques attribuables aux différentes IDG. Il est aussi possible de voir les thématiques en commun entre deux IDG, permettant ainsi de connaître l'intersection de leurs couvertures sémantiques. Image provenant de [62].*

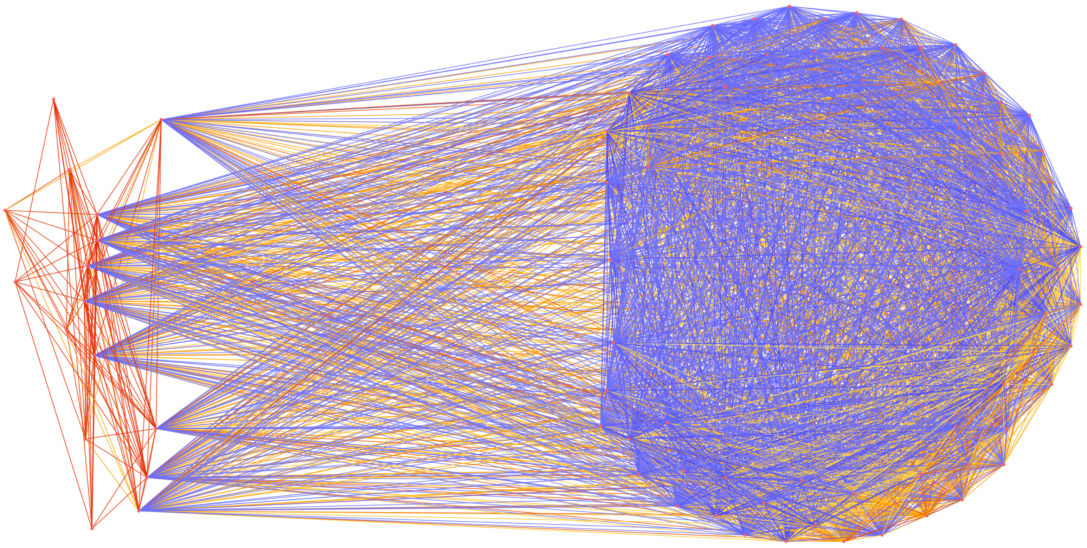

FIGURE *2.4* – ***Graphe d'emprises géographiques*** *: Extrait du graphe d'emprise de PIGMA (l'IDG Aquitaine). Les sommets représentent les études géographiques et une arête entre deux sommets indique que les emprises spatiales des deux études se superposent et que leur surface d'intersection est supérieure à un seuil fixé par les experts. Plus la surface d'intersection est élevée plus l'arête tend du bleu vers le rouge. Dans ce graphe, on peut ainsi voir que l'essentiel des études ne se recouvrent que légèrement (arêtes bleues), signe d'une couverture spatiale homogène, exceptée une communauté d'études hautement superposées (arêtes rouges).*

nous avons disposé sont entièrement physiques : un corpus composé de plusieurs dossiers judiciaires d'environ 25000 pages chacun, traitant d'affaires répertoriées par la police et relatives à ces réseaux. De nombreux types de documents y sont consignés : témoignages, écoutes téléphoniques, rapports d'interrogatoire, liste de numéros de téléphones suspects, rapports de police...

Une étape essentielle va donc être de numériser ces données afin de pouvoir les visualiser et les exploiter efficacement. Cette étape est difficilement automatisable notamment à cause de l'hétérogénéité des documents et de la mauvaise qualité d'impression des feuilles excluant de scanner automatiquement les dossiers. En plus de cela, beaucoup de document nécessitent une analyse humaine afin de s'assurer de la pertinence voir de la véracité des informations. Un interview, un témoignage ou un interrogatoire peut apporter des informations en contradiction avec d'autres, non fiables ou même volontairement fausses (exemple d'une personne ne donnant

pas son vrai nom, numéro de téléphone, etc.). Ces éléments requièrent alors une intervention humaine mais, avec des dizaines de milliers de page à étudier, il est nécessaire d'avoir une aide informatique afin de stocker et interroger efficacement ces données.

Pour ce faire, un modèle abstrait de données a été réalisé et utilisé au sein d'une plateforme en ligne permettant de faciliter la consultation et la saisie des informations. Cette plateforme a été réalisée simultanément avec l'exploration des dossiers par nos collègues des sciences humaines et sociales, le modèle abstrait a donc évolué au fur et à mesure des découvertes et a été mis à jour régulièrement en même temps que la plateforme était implémentée.

Pour cette raison, nous avions commencé le projet avec une base de données relationnelle, solution classique pour des données basées sur des relations. Les données permettent de définir un réseau multi-couche (Fig. 2.5) où chaque couche est définie par les types de relations : lien financier, lien sexuel, lien de sang, lien de réseau, lien de connaissance, lien de soutien et lien juju (une cérémonie religieuse incitant une personne à se prostituer pour rembourser une dette sous peine de "mauvais sort" [46]).

Cependant, cette solution s'est vite révélée problématique au niveau de la conception des requêtes et de leurs performances. Les requêtes extrêmement complexes, en raison du grand nombre de jointures dû aux nombreux types de liens et entités, ainsi que la nécessité de changer ou faire évoluer régulièrement le modèle de données ne conviennent pas à la rigidité du modèle relationnel [45].

C'est pourquoi nous avons, dans le cadre du CVCE, stocké et utilisé ces données à travers, d'une part, une base de donnée graphe (une base de données spécialement conçue pour l'exploitation des réseaux) et, d'autre part, Tulip [3], une infrastructure logicielle spécialisée dans les réseaux afin de bénéficier d'une souplesse dans la conception du modèle (et ainsi permettre son évolution) ainsi qu'une performance accrue pour toute requête nécessaire à l'analyse ou la visualisation. Plus de détails sur le rôle et l'utilisation de ces objets sont disponibles à la section 6.1.

2.4 Synthèse

A travers les objectifs et les données de nos collègues historiens du CVCE, nous proposons une méthode permettant à la fois une navigation simple et intuitive dans leur corpus de documents ainsi qu'un respect facilité des contraintes inhérentes à la conception des ePublications.

Pour ce faire, les données dont nous disposons prennent la forme d'une interconnexion complexe de sous-systèmes : un réseau multi-couche. Cet objet nécessite cependant des techniques et méthodes spécialisées pour pouvoir bénéficier pleinement de sa sémantique riche, que nos expériences passées ont fait émerger notamment à travers **une différenciation des traitements des couches**, **un focus accru sur les éléments intéressants pour l'utilisateur** et **des choix techniques optimisés pour ce type de réseau**.

Cependant, concevoir une méthode de navigation et de visualisation pour des sciences informatiques ou physiques n'est pas le même exercice que pour les sciences humaines et sociales. Les méthodologies sont susceptibles de différer, tant par les types de visualisation nécessaires que la manière d'utiliser ces visualisations.

Dans le chapitre suivant, nous faisons donc un tour d'horizon des spécificités des sciences humaines et sociales afin de déterminer les éléments et contraintes nécessitant d'être pris en compte pour modeler notre méthode au plus près des besoins et méthodologies de ce domaine.

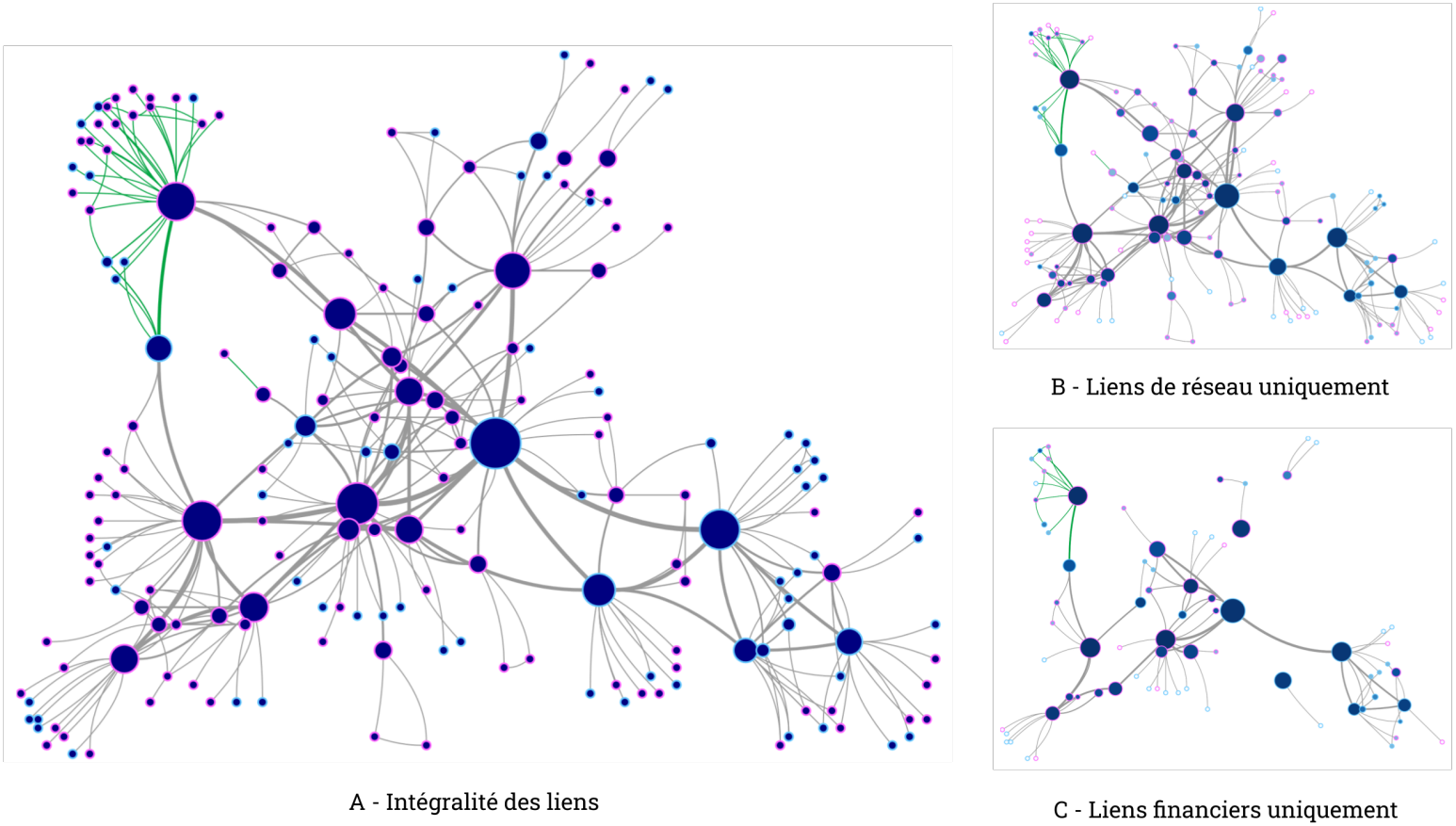

FIGURE *2.5* – ***Réseau TETRUM*** *: Voici un extrait du réseau TETRUM. Chaque sommet représente une personne dont la couleur du contour indique le sexe. Le réseau A correspond au réseau complet. Les réseaux B et C représentent respectivement uniquement les liens de réseau (témoigne d'une action au sein du réseau criminel entre deux personnes) et les liens financiers (témoigne d'un échange de nature financière entre deux personnes) ainsi que les sommets reliés par ces liens. Les liens verts correspondent à la même sélection de sommets dans les trois réseaux présentés.*

Visualisation et sciences humaines et sociales

Sommaire

La navigation et l'exploration de réseaux est à la croisée des chemins de plusieurs domaines. Il est évidemment nécessaire de s'immerger dans l'analyse structurelle et sémantique des réseaux afin de comprendre les objets que nous utilisons et s'adapter en conséquence (c'est ce qui a été réalisé dans le chapitre 2) mais il est tout autant nécessaire de se focaliser sur la visualisation générale en elle même.

Si jusqu'ici nous avons toujours traité d'éléments relatifs aux graphes et aux réseaux, la visualisation couvre cependant un spectre d'objets bien plus large [67]. Plus exactement : là où il y a des données, il y a de la visualisation.

Ce domaine n'est pas nouveau. Déjà dans les années 1850, Charles Joseph Minard a produit plusieurs représentations graphiques de données et peut être considéré comme un pionnier de la visualisation analytique. Dans plusieurs de ses travaux (notamment ses "cartes figuratives et approximatives" : "des quantités de viandes de boucherie envoyées sur pied par les départements et consommateurs à Paris", "des tonnages des Grand Ports et des principales Rivières d'Europe", "représentant pour l'année 1858 les émigrants du globe", etc. - voir Annexe A), C.J Minard essaye déjà d'exprimer visuellement les interconnexions entre de multiples

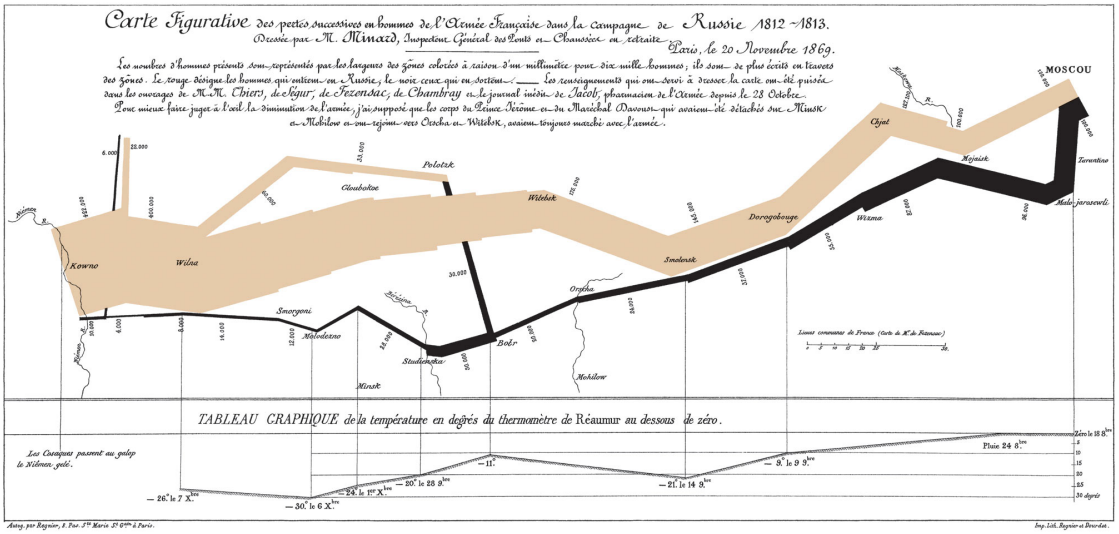

FIGURE *3.1* – ***Début de la visualisation*** *: Ceci est le travail de Charles Joseph Minard, un des pionniers de la représentation graphique des données. Cette carte réalisée en 1879 est son travail le plus célèbre et représente le flux des soldats français lors de la campagne de Russie en 1812–1813 sous le règne de Napoléon. La largeur du trait orange représente le nombre d'individus encore en vie et le trait noir le nombre de pertes humaines. Ces éléments peuvent ensuite être analysé par rapport à l'emplacement géographique, la température et les divers renforts indiqués par les traits transversaux.*

types de données. Son travail le plus célèbre, la "Carte figurative des pertes successives en hommes de l'Armée française dans la campagne de Russie en 1812–1813" (Figure 3.1, https://fr.wikipedia.org/wiki/Carte_figurative), détaille simultanément de nombreuses informations différentes : la chronologie de l'évènement, la température, l'itinéraire des soldats, la position géographique, le nombre de soldats à travers le temps ainsi que les renforts et pertes humaines.

Ces travaux sont précurseurs d'une des problématiques de la visualisation moderne : comment présenter un ensemble de données de différents types simultanément et efficacement. S'il est trivial de recouper deux types de données afin d'analyser les relations qu'ils entretiennent, le problème se complexifie lorsque de nouvelles dimensions sont ajoutées aux données étudiées. Cette problématique, présente depuis longtemps tant en science des données (avec par exemple les questionnements relatifs à l'implémentation de ces visualisations plus complexes [79]) qu'en sciences humaines (avec par exemple les travaux menés sur des données à la fois spatiales

et temporelles [34] avec les 'Space-time cubes", revisités de nombreuses années plus tard [42] et qui forment encore aujourd'hui la base de la "time-geography"), prend une importance capitale avec l'étude des graphes multi-couches.

Les réseaux multi-couches (voir Chapitre 2) font partis de ces objets complexes, aux multiples dimensions, qui nécessitent des méthodes spécifiques pour exprimer leur plein potentiel. S'il est toujours possible d'occulter des caractéristiques d'un réseau multi-couche pour le réduire à un objet facilement visualisable (souvent un graphe simple), beaucoup de domaines, dont les sciences humaines, ont un intérêt à conserver l'entièreté des informations. De ce fait, des méthodes spécifiques doivent être déployées. C'est ce que nous proposons de faire avec `M-QuBE`^{3}, une méthode d'exploration spécialisée pour les réseaux multi-couches dans le cadre des sciences humaines et sociales conçue afin de répondre aux besoins de nos collègues historiens et s'appuyant sur les expériences accumulées auprès de collègues juristes, sociologues et géographes.

Dans ce chapitre, nous faisons un tour d'horizon du domaine de la visualisation (section 3.1) puis de quelques unes de ses applications dans le domaine des sciences humaines et sociales. A travers les liens avec la science des données (section 3.2), les réseaux multi-couches (section 3.3) et les méthodes exploratoires (section 3.4), nous explicitons les spécificités des sciences humaines et sociales afin de justifier les choix qui ont été faits dans la conception et le développement de `M-QuBE`^{3} décrit dans les chapitres suivants.

3.1 Visualisation : méthodes et mantras

Comme dit précédemment, la visualisation n'est pas un concept récent. Ce domaine s'est construit graduellement en suivant des préceptes dont certains ont été hissés au rang de mantra. Intervenant à deux niveaux, ces règles sont liées d'une part à la conception de la visualisation et d'autre part à l'application même de la visualisation.

Le premier et plus connu des mantras est celui de Shneiderman [70] qui définit une caractérisation et l'ordre des interactions nécessaires : **"Overview first, zoom and filter, then details-on-demand"** (on établit d'abord une vue d'ensemble, ensuite on zoom et on filtre puis on détaille à la demande). Par ce mantra, on détermine une méthodologie à adopter pour maximiser l'efficacité de la visualisation. Ceci impacte lourdement la phase de conception car nécessite des interactions et des visualisations spécifiques. En effet, une processus de visualisation n'est pas nécessairement un processus fixe. Cela peut à la fois désigner un processus à une

seule étape, où la visualisation va être calculée puis présentée comme conclusion à l'utilisateur, ou suivre une méthodologie en plusieurs étapes, suivant par exemple le mantra précédemment cité, permettant de rentrer dans les données, d'y naviguer et d'en faire ressortir ce qui est intéressant pour un utilisateur donné [78]. Ce second scénario correspond à la notion d'exploration, étroitement corrélée à la visualisation tant l'exploration de données ne peut s'effectuer sans visualisation et que la visualisation ne se justifie parfois que par l'exploration.

Le domaine de l'exploration de données devient progressivement indispensable à l'heure où nos usages quotidiens d'internet, des objets connectés ou des réseaux sociaux génèrent des quantités phénoménales de données. Comme dit précédemment, exploration et visualisation sont liées et suivent des règles communes. Cependant, le domaine d'application qui nous intéresse dans le cadre de notre projet, i.e. celui des sciences humaines et sociales, a des spécificités qui influent sur la méthodologie habituelle.

En effet, le mantra de Shneirderman est habituellement amplement suffisant pour mener à terme une visualisation efficace sur les données : à partir d'une vue globale, l'utilisateur peut cibler les points d'intérêt ou, en cas de difficulté, filtrer cette vue pour les trouver plus aisément. Puis, on concentre l'analyse sur ces points. Cependant, si les données sont trop volumineuses ou connectées, une vue d'ensemble naïve posera des problèmes de lisibilité à cause d'un encombrement visuel trop important [30, 31] et ne répondra donc pas à son objectif. Un solution communément utilisée est alors d'agréger les données afin d'offrir une vue d'ensemble simplifiée mais significative. L'agrégation nécessite alors de pouvoir grouper des informations pour conserver et mettre en avant essentiellement ce qui est porteur de sens et d'intérêt pour l'utilisateur. Voici quelques exemples représentatifs nécessitant l'agrégation des données :

- imMens [50] est un système de visualisation dont les méthodes utilisées sont implémentées et optimisées pour fonctionner sur les navigateurs internet. Le credo autour d'imMens est de proposer des visualisations dont la limite ne serait pas le volume de données à visualiser mais la résolution de l'écran utilisé pour la visualisation. Un tel objectif induit forcément de faire des concessions sur l'affichage en déterminant précisément ce qui sera montré à l'utilisateur pour éviter une surcharge visuelle qui détériorerait la lisibilité. De plus, l'interactivité et l'exécution en temps réel de la visualisation ajoutent une couche de difficulté supplémentaire et nécessitent encore davantage à agréger les données visualisées pour conserver des performances stables. Il y a donc nécessité de représenter comme des éléments individuels des groupes

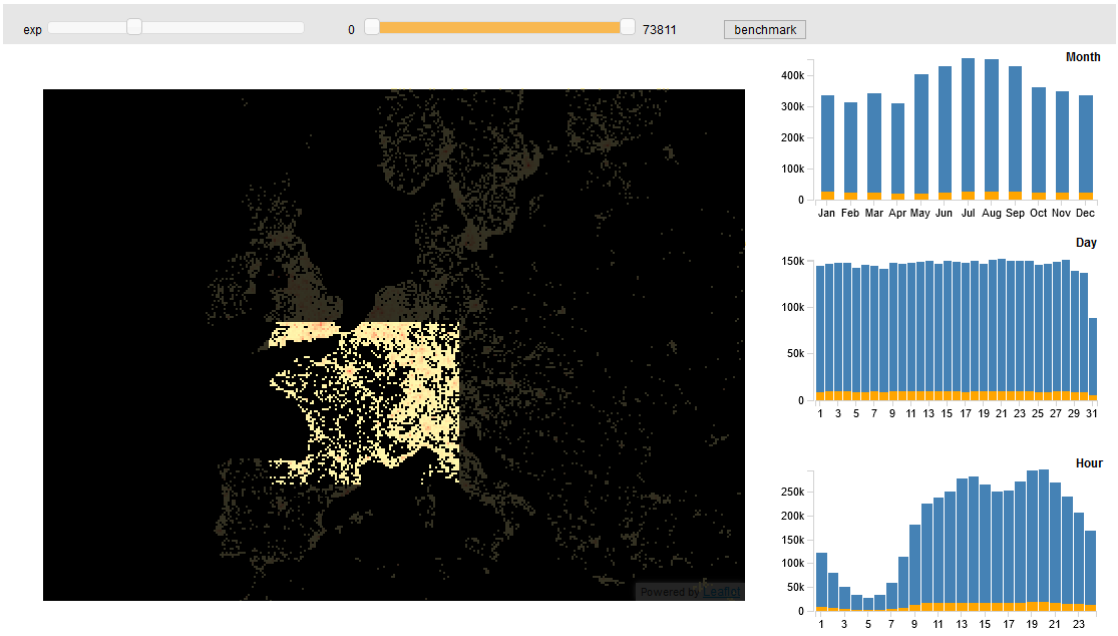

FIGURE *3.2* – ***imMens*** *: Ceci est une capture d'écran provenant d'une version de démonstration d'imMens. La vue de gauche est une carte de chaleur représentant le nombre de personnes enregistrées pour des voyages en avion sur une période d'un an. Les trois autres visualisations montrent ces même enregistrements en fonction des mois, des jours et des heures. En sélectionnant une zone comprenant la France sur la carte (points jaunes), on peut observer l'ensemble des éléments correspondant dans les trois autres visualisations. Ainsi sélectionner une zone de quelques pixels sur la carte peut sélectionner plusieurs milliers d'éléments dans les autres vues.*

d'éléments ou des ensembles d'informations : sélectionner un point sur imMens correspond alors à sélectionner un ensemble potentiellement très grand de valeurs liées à ce point (voir Figure 3.2).

- Si imMens propose des visualisations sur des données non inter-connectées, il existe aussi de nombreux exemples pour la visualisation de graphes et réseaux. C'est par exemple le cas de JASPER [73] qui se propose de représenter un réseau entier sous forme de mosaïque pixelisée (voir Figure 3.3) dont les carreaux (les zones d'une même couleur) représentent les communautés. Les pixels de ces zones représentent les différents éléments constituant les communautés qu'elles définissent. Une telle visualisation fait le choix d'orienter la

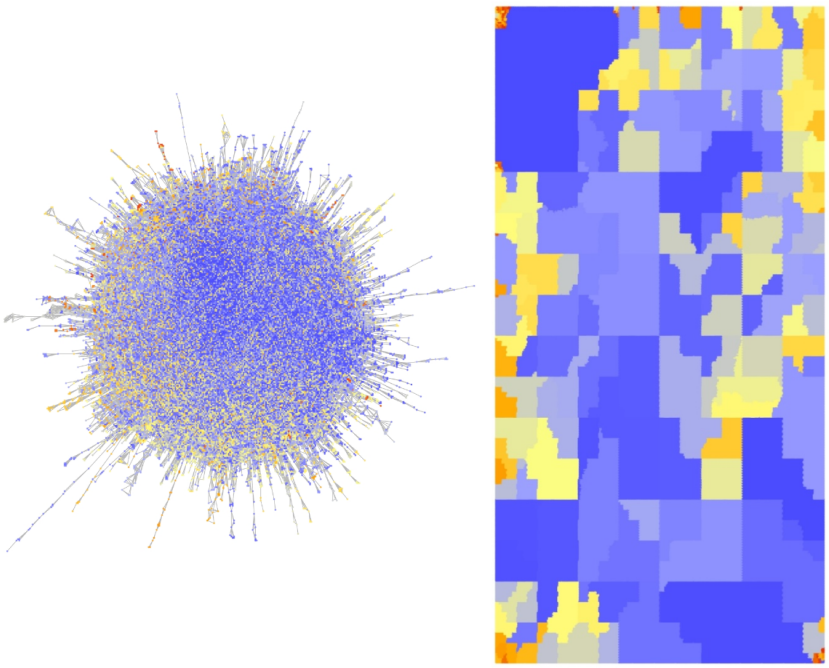

FIGURE *3.3* – ***JASPER*** *: Cette méthode propose une représentation orientée pixel permettant de mettre en avant les communautés et leur disposition au sein du réseau en agrégeant les données représentées. La visualisation de gauche correspond à la représentation noeud-lien des données (utilisant l'algorithme de dessin* FM^3 *[33] pour déterminer la position des noeuds). Celle de droite représente les même données mais utilisant JASPER. Les zones de couleurs représentent les différentes communautés et les pixels les composants sont les différents noeuds du réseau. Image provenant de [73].*

visualisation pour se concentrer essentiellement sur les communautés, quitte à ne pas représenter les liens entre les différents sommets. Ces liens sont néanmoins pris en compte afin de conserver une proximité entre les sommets voisins dans la visualisation. Cela permet ainsi d'avoir aisément un aperçu représentatif des relations inter-communautaires, de l'étendue de ces communautés ou de l'évolution du réseau lors d'une étude de propagation.

- Un autre exemple similaire sont les travaux autour de la navigation multi-échelle [4,12]. Le concept est de transformer un graphe dont la taille et la complexité ne permettent pas une lecture simple à un objet plus réduit. Ce faisant, des structures visuelles autrement imperceptibles apparaissent, rendant

le graphe plus facilement analysable. De nombreuses méthodes existent [49] passant par exemple par le groupement d'arêtes (représentation d'un ensemble d'arêtes du graphe en une seule arête) ou la création de méta-noeud (représentation d'un sous-réseau par un unique sommet). Ceci nécessite néanmoins de pouvoir hiérarchiser les données afin de déterminer ce qui est agrégable, ce qui n'est pas toujours possible.

Si le mantra de Shneiderman est le plus couramment utilisé, il n'est pas nécessairement adapté à tous les scénarios. Celui-ci est essentiellement centré sur la visualisation. Elle est le point de départ du processus et l'entièreté de l'analyse en dépend. Une approche davantage basée sur l'analyse visuelle a été proposée par Keim *et al.* dans un nouveau mantra inspiré de celui de Shneiderman : **"Analyze first, Show the Important, Zoom, filter and analyze further, Details on demand"** [39] (d'abord on analyse, ensuite on montre ce qui est important, on zoom, on filtre, on analyse plus en profondeur puis on détaille à la demande). Celui-ci ne s'abstrait pas de la nécessité d'utiliser un vue globale simplifiée. Les deux premières étapes : l'analyse et la mise en valeur de ce qui est intéressant nécessite une agrégation des données pour procéder aux phases postérieures de la visualisation (avec le zoom, le second filtrage et l'éventuelle extension de ce qui est montré). Cela peut être contraignant voir rédhibitoire si une vue globale est difficile ou semble impertinente dans le scénario (voir section suivante).

Chaque nouveau scénario apporte son lot de variations et de contraintes susceptibles d'imposer un ajustement des mantras existants. On peut néanmoins classer les différents travaux vus précédemment en deux catégories :

- **Top-Down :** Comme son nom l'indique, les visualisations de type *top-down* sont celles qui vont commencer par établir une vue générale (*top*) avant d'offrir un moyen de se pencher à posteriori sur les détails des données (*down*). Toutes les visualisations citées en exemple dans cette section sont de type *top-down* car découlant du mantra de Shneiderman ou Keim. Elles sont contraintes de passer par de l'agrégation, du filtrage ou de s'appliquer à des jeux de données qui sont pas trop volumineux ou complexes.
- **Bottom-Up :** Inversement, l'approche *bottom-up* propose de partir au plus près des données (*bottom*) puis, à partir de ces données, de pouvoir analyser ou se faire un aperçu de l'ensemble (*up*). Cette approche est privilégiée lorsqu'une vue générale n'est pas possible à cause des contraintes liées au scénario ou au domaine de l'étude. C'est par exemple le cas pour la visualisation d'ontologie [23] qui n'a pour le moment pas moyen de mener une

étude sans se focaliser en premier lieu sur des éléments précis et ne peut donc pas se baser essentiellement sur une vue globale. Une illustration de cette approche sont les travaux de S.van den Elzen et J.J. van Wijk [75]. Mêlant simultanément une vue centrée sur les détails et une vue globale des données au sein d'une même visualisation, ils proposent par une analyse et une sélection d'éléments précis de générer une vue agrégée en fonction des éléments intéressants pour l'utilisateur. Cette approche impose néanmoins aussi des contraintes : il est nécessaire de pouvoir naviguer dans le détail des données efficacement, un mécanisme de filtrage et de sélection est donc absolument nécessaire. Plus qu'en science des données, cette approche devient incontournable pour certains domaines dont les sciences humaines et sociales, ce que nous verrons dans la section suivante.

Nous avons présenté les approches classiques de la visualisation. Cependant, la structure multi-couche de nos données (cf. Chapitre 2) et les contraintes inhérentes aux sciences humaines et sociales nous obligent à diverger des méthodes traditionnelles. Dans la partie suivante, nous détaillons ces spécificités et ce qu'elles induisent afin d'expliquer, d'une part, les challenges relevés par notre méthode et, d'autre part, les particularités de notre méthode.

3.2 Science des données et sciences humaines et sociales

La science des données et les sciences humaines et sociales ont des manières distinctes d'initier et de mener leurs analyses respectives. Par leurs différences d'objectifs et de données, les méthodologies sont inévitablement différentes. Ces particularités et cette divergence par rapport à la science des données sont notamment développées dans les travaux de Borgatti et al. [9]. L'analyse et l'exploitation des réseaux sont ancrées dans un large panel de domaines des sciences humaines, de l'Histoire à l'économie en passant par la psychologie. La différence fondamentale avec la science des données est que l'attention des experts va se concentrer non pas sur une analyse du réseau lui même mais sur "comment des individus autonomes peuvent se combiner pour créer des sociétés durables et fonctionnelles". L'essentiel de l'attention est alors placé dans une analyse rapprochée des éléments du réseau. Borgatti cite en exemple les travaux de Moreno [59] qui fait l'analogie entre ces réseaux et des systèmes physiques. Les personnes et autres éléments constitutifs du réseau analysé deviennent alors des "atomes sociaux" soumis aux lois de la

“gravitation sociale” (lois impactant l’apparition des liens et donc la structure du réseau), faisant ainsi écho au terme de “physique sociale” du philosophe français Auguste Comte [18] presque cent ans plus tôt. Pour accentuer encore davantage l’importance capitale mise sur les individus, il cite aussi l’ouvrage de Durkheim [25] où les sociétés humaines sont comparées à des organismes biologiques constitués de l’interrelation de leurs éléments. On pourrait citer aussi la notion d’“ego network” [2, 55], au nom explicite, où l’analyse s’effectue à partir d’un individu (*ego*) pour analyser les différentes couches de son entourage (*alter*), toujours afin de centrer l’analyse sur l’individu et son entourage. On observe alors deux niveaux d’analyse, une niveau inter-individuel, se concentrant sur les individus, et un niveau inter-organisationnel, se concentrant sur les communautés d’individus. Si ces deux perspectives se voient accorder une importance différente en fonction du courant dans lequel se situent les experts (individualisme ou holisme), elles ne sont pas nécessairement distinctes et il est possible de les prendre en compte conjointement afin d’établir une étude plus globale [48]. La notion de “Noyau-Périphérie” [8], où l’on considère un groupe indivisible d’individus (*core*) connectés à des individus externes (*periphery*), est déjà positionnable de manière intermédiaire entre le niveau inter-individuel et le niveau inter-organisationnel. Cette notion est même élargie à un modèle continu, où il existe potentiellement plusieurs classes d’individus semi-périphériques, amenuisant encore plus les différences entre ces deux niveaux.

Parmi tous ces modèles, le point commun est l’importance apportée à ce qui constitue le réseau plutôt que le réseau lui même. C’est notamment ce que dit Borgatti lorsqu’il développe les perceptions mutuelles entre la science des données et les sciences sociales [9]. Chaque domaine est susceptible de considérer l’autre comme uniquement descriptif. Les scientifiques peuvent reprocher aux sociologues de ne pas opposer leurs mesures des propriétés du réseau à des modèles théoriques là où des sociologues peuvent considérer les travaux des scientifiques comme horriblement simplistes et génériques (“*alarmingly simplistic and coarse-grained*”). Borgatti indique par exemple que des modèles d’étude comme ceux basés sur l’utilisation de graphes aléatoires paraissent extrêmement naïfs pour les sociologues, s’apparentant pour eux à “comparer un gratte-ciel à une distribution aléatoire de la même quantité de matériaux”. Borgatti explique cette différence en sciences sociales et humaines par un intérêt supérieur pour le sommet individuel (représentant tant un individu qu’un collectif d’individus) que pour le réseau lui même en tant qu’ensemble.

Il s’agit donc essentiellement d’un paradigme différent auquel notre méthode va devoir s’adapter pour être en accord avec la méthodologie et les objectifs généraux

des sciences humaines et sociales. En plus d'opter pour une approche majoritairement centrée sur les sommets, il est aussi nécessaire de capturer et d'exploiter le caractère multi-couche des réseaux utilisés en sciences humaines et sociales. C'est ce que nous proposons dans la section suivante.

3.3 Réseaux multi-couches et sciences humaines et sociales

En sciences humaines et sociales, l'étude d'un réseau n'est pas tant l'analyse du réseau lui même (topologie) que directement l'analyse de la sémantique qu'il véhicule. Or, comme dit précédemment (voir Chapitre 2), un réseau sémantiquement riche est facilement modélisable sous forme de réseau multi-couche.

Avant même que le concept de multi-couche se démocratise, ces objets étaient déjà traités inconsciemment même sans avoir spécifiquement de méthodologies affirmées centrées sur les réseaux multi-couches. Nos expériences, tant avec les géographes de GEOBS que les juristes et sociologues de TETRUM (voir Chapitre 2), reflètent d'ailleurs cette situation. Les géographes ont par exemple souvent cherché à générer des graphes bipartis en utilisant deux types parmi l'ensemble des métadonnées dont on disposait [62]. La situation est alors similaire au fait d'établir un filtrage sur le réseau multi-couche correspondant à l'ensemble des données pour ne conserver et analyser que les interactions de deux de ses couches. Pour TETRUM, la conception et le développement du modèle de données a mené progressivement à considérer notre réseau comme multi-couche [45]. Si les sommets du réseau criminel représentent tous des personnes, le caractère multi-couche du réseau vient cette fois du large spectre de types de lien possibles entre ces personnes (liens financiers, liens de prostitutions, liens de sang, etc.). On parle alors de réseau multiplexe, une sous-catégorie de réseau multi-couche où les couches sont définies non pas par le type des sommets mais par le type des liens du réseau [41]. D'autres exemples de travaux sur les réseaux criminels [13, 68] comportent aussi cette caractéristique : même sans expliciter l'aspect multi-couche du réseau, l'analyse est menée en ayant une pleine connaissance de ses couches voir en concentrant l'analyse sur celles-ci.

Pour autant, il est courant de voir en sciences humaines et sociales des réseaux exploités comme des graphes simples. C'est le cas par exemple des travaux précédemment cités de Borgatti [8] sur la notion de "Noyau-Périphérie". Dans cet exemple, les sommets dit noyaux (groupes sémantiquement indivisibles d'éléments), ceux appartenant à la périphérie ainsi que ceux appartenant aux différents

niveaux intermédiaires (pour le modèle continu) sont le résultat d'une analyse structurelle du réseau ayant pour but de faire ressortir des classes d'éléments sémantiquement significatifs. Ces classes sont néanmoins estimées indépendamment des types d'origine des sommets alors qu'ils pourraient pourtant avoir une importance élevée lors d'une analyse sémantique du réseau. D'une manière plus générale, les études basées sur du partitionnement [35,58] et autres analyses structurelles [54] ne différencient pas nécessairement les types de sommets et exploitent donc le réseau multi-couche comme sil était mono-couche.

Il existe néanmoins des cas d'utilisation du caractère multiplexe des réseaux. C'est par exemple le cas des travaux de Perer et Shneiderman [65]. Ceux-ci y proposent un système de visualisation permettant de trier et filtrer les sommets en fonction de métriques (degrés, différents types de centralité, barycentres, etc.) afin d'obtenir une liste ordonnée de sommets. Ceux-ci sont ensuite affichés à travers une vue noeud-lien personnalisable où l'utilisateur peut décider d'afficher uniquement certaines couches (donc uniquement les sommets ayant un certain type de lien) ainsi que des sommets ayant un certain classement ou score via les métriques. Le caractère multi-couche du réseau est ici utilisé mais comme filtre visuel et non pour établir la pertinence des sommets à afficher.

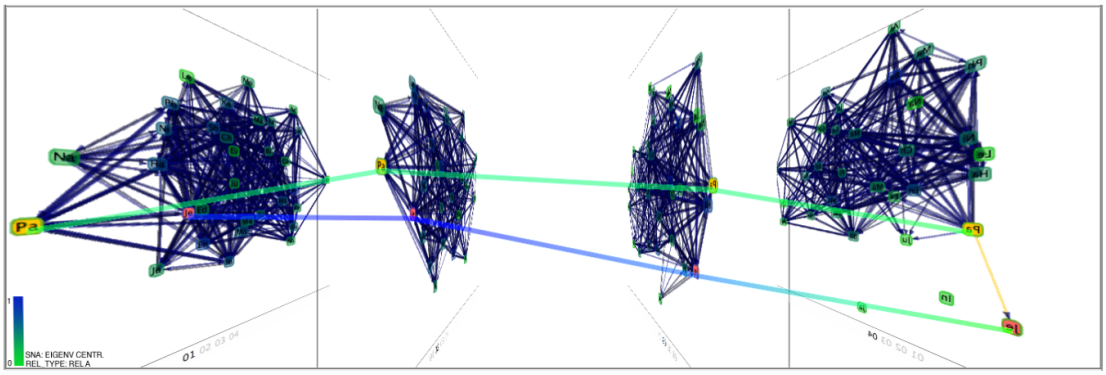

FIGURE *3.4* – ***Visualisation en 2.5D*** *: Ceci est une visualisation en 2,5D de quatre ensembles de sommets d'un réseau dynamique. Chacun représente une couche temporelle différente. Les liens entre les différentes couches permettent de suivre l'évolution des sommets à travers les différentes périodes de temps. Image provenant de [26].*

Un autre exemple ne référençant pas directement l'aspect multi-couche mais pouvant néanmoins l'utiliser est l'approche de Federico et al. [26] sur les réseaux

dynamiques. Dans ces travaux, une visualisation en 2.5D (Figure 3.4) représente plusieurs ensembles de sommets correspondant chacun à des périodes de temps. Les liens entre ces ensembles de sommets donnent alors une indication de l'évolution temporelle des sommets en suivant leurs présences dans les différents ensembles. On peut concevoir ces ensembles comme autant de couches dont le type est déterminé par leurs dates. Ce faisant, on peut établir une visualisation des couches et des liens inter-couches d'un réseau. Une adaptation aux graphes multi-couches de la visualisation en 2,5D a d'ailleurs été réalisée par un partenaire du projet BLIZAAR dont un aperçu est accessible sur le site du projet (https://blizaar.list.lu/doku.php?id=eisti_tool).

Comme nous venons de voir, plusieurs exemples exploitant plus ou moins étroitement le coté multi-couche des réseaux sociaux existent. Cependant, notre méthode a besoin de pouvoir utiliser directement cette particularité pour impacter les visualisations. La solution adoptée doit donc d'une part pouvoir être spécialisée dans l'analyse de réseau et d'autre part prendre en compte tous les types (et donc couches) des sommets pour définir une visualisation qui sera pertinente pour l'utilisateur.

Enfin, il reste à déterminer comment explorer ces réseaux. Notre méthode doit donc utiliser les particularités des sciences humaines et sociales afin de définir une manière de naviguer et d'interagir permettant une exploration efficace. C'est ce que nous détaillons dans la partie suivante.

3.4 Exploration et sciences humaines et sociales

L'exploration est le but initial de nos travaux. Nos collègues ayant de vastes jeux de données, il est nécessaire de mettre au point un mécanisme permettant de naviguer efficacement et sans se perdre dans les données. S'il existe déjà des outils et méthodes pour visualiser des réseaux multi-couches [21,22,56], il y a néanmoins peu de méthodes spécialisées à la fois pour l'exploration et les sciences humaines et sociales comme l'attestent Perer et Shneiderman en conclusion de leurs travaux précédemment cités [65]. Pour répondre à ce problème et mettre au point une méthode adaptée aux besoins des experts des données, il est nécessaire de concevoir notre méthode en prenant en compte tant les spécificités du domaine que leur méthodologie. Nous avons vu précédemment qu'une attention toute particulière est placée à l'échelle de l'individu ou du groupe d'individus plutôt que sur le réseau lui même. L'exploration, comme l'analyse, se doit donc de suivre ce schéma est de pouvoir être menée directement à partir des éléments du réseau et de leurs

détails. Ceci fait écho à une approche que nous avons vue dans une des parties précédentes (Section 3.1), le *bottom-up*.

Les approches *bottom-up* permettent d'initier analyses et explorations au plus près des données afin de pouvoir à posteriori élargir le spectre d'analyse à des communautés, sous-réseau voir réseau tout entier. Les travaux de S.van den Elzen et J.J. van Wijk [75], précédemment cités, s'ancrent dans cette dynamique en permettant la génération d'une vue agrégée de l'ensemble des données en fonction des sélections effectuées à l'échelle des individus. Ainsi, en prenant connaissance des éléments et de leurs interconnexions à l'échelle atomique, l'utilisateur va forger par ses interactions une vue plus globale permettant de comprendre la structure du réseau à l'échelle globale. Un autre exemple plus amplement lié aux sciences humaines et sociales est l'étude menée par Ghanie et al [29]. Le cadre de ces travaux est très similaire au notre : une coopération avec des experts en sciences humaines et sociales où les données sont modélisables sous forme de réseau multicouche. Une nécessité de simplification visuelle est également nécessaire afin de pouvoir créer des visualisations utiles et exploitables par les experts des données. Parmi les différentes alternatives présentées afin de réduire la complexité visuelle, se trouve notamment le "diviser pour mieux régner" (*divide and conquer*) : subdiviser le problème ou les données à visualiser en sous-ensembles de tailles réduites jusqu'à obtenir un résultat compréhensible et analysable. Ce concept synergise particulièrement bien avec les sciences humaines et sociales en répondant d'une part au problème de complexité visuelle et en permettant d'autre part d'axer la visualisation sur une échelle proche des individus et groupes d'individus.

La méthode utilisée, PNLBs ("*Parallel Node-Link bands*"), est d'ailleurs similaire dans son concept aux travaux menés par notre collègue du projet BLIZAAR (dont un aperçu est disponible sur le wiki officiel : `https://blizaar.list.lu/doku.php?id=list_tool`). Le réseau est présenté à l'expert sous la forme de listes de sommets inter-connectées entre elles à la manière de coordonnées parallèles où les listes représentent les différentes couches du réseau. Encore une fois, la visualisation est essentiellement menée à l'échelle des sommets et permet de fait de montrer une différence au niveau des paradigmes habituellement utilisés. C'est une des particularités de l'approche *bottom-up*, les mantras habituels tel que le "Overview first, zoom and filter, then details-on-demand" de Shneiderman [70] ou le "Analyze first, Show the Important, Zoom, filter and analyze further, Details on demand" de Keim [39] ne sont pas adaptés. Plus que de simples différences de méthodes, le *bottom-up* induit un changement au niveau même du paradigme utilisé.

Un certain nombre de travaux ont d'ailleurs proposé leurs propres versions

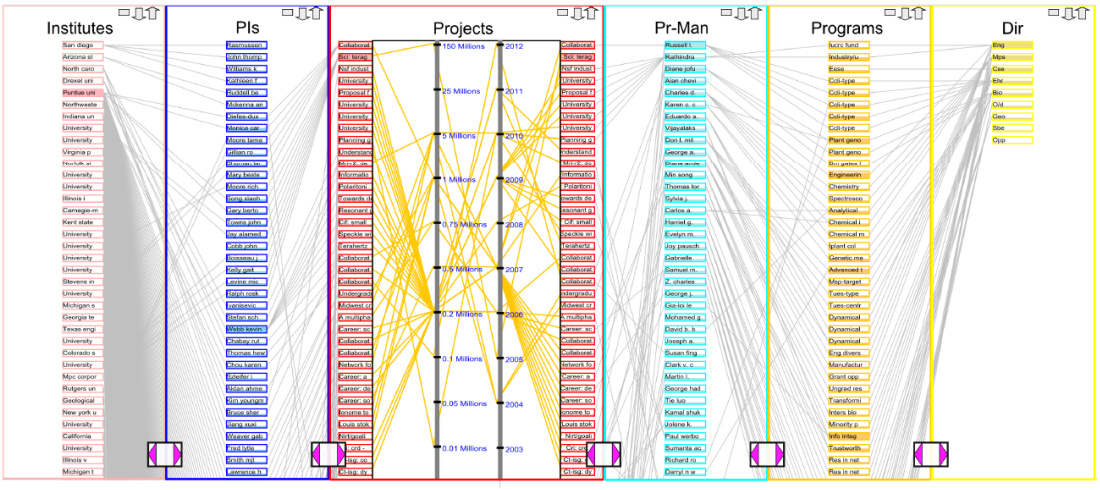

FIGURE *3.5* – ***PNLBs : visualisation par listes***. *Cette visualisation présente les couches du réseau sous forme de listes inter-connectées, mettant ainsi l'accent sur les connexions inter-couches. Image provenant de [29].*

des mantras de visualisations ré-adaptés au *bottom-up* comme le "Details-first, show context, overview last" de Luciani et al [52] (détailler en premier, montrer le contexte et finir par une vue générale) ou le "Search, show context, expand on demand" (chercher, montrer du contexte, étendre à la demande) de Van Ham et Perer [76] dont nous parlerons plus en détail dans le chapitre 5. Nos travaux s'inscrivent dans ce dernier par la proximité avec les habitudes des experts : à partir d'un point de départ connu (ex : une personnalité politique, un évènement historique, une communauté sociale donnée, etc.), on enrichit sémantiquement le contexte (on met en valeur les éléments environnants, les connexions à des groupes significatifs, etc.) puis on étend l'analyse si nécessaire (à des pistes connexes, transversales ou simplement en enrichissant sémantiquement encore davantage le contexte actuel).

Si le *bottom-up* est parfaitement en accord avec le fondement des sciences humaines et sociales, il reste néanmoins un aspect de leur méthodologie qui doit être impérativement pris en compte. L'exploration en sciences humaines et sociales prend souvent la forme d'un tâtonnement : on part d'un élément défini sans savoir précisément à l'avance vers où notre piste va nous mener. Dans certains cas, une piste peut mener vers d'autres pistes et le processus d'exploration est alors à recommencer à partir d'un nouveau référentiel pour explorer ces nouvelles pistes, et ainsi de suite... La méthode d'exploration doit donc prendre en compte une dé-

marche itérative où l'on doit pouvoir essayer et ré-essayer de visualiser et explorer différentes pistes voir même prendre en compte l'aspect arborescent d'une telle procédure, chaque piste pouvant mener à d'autres pistes.

3.5 Vers la méthode M-QuBE^3

A notre connaissance, une méthode proposant à la fois une **échelle centrée autour de l'individu**, **une gestion du caractère multi-couche des réseaux utilisés en sciences humaines et sociales** et **une gestion de l'aspect itératif et arborescent de la méthode de travail du domaine**, n'a pas encore été développée. Pour répondre à ce besoin, nous proposons notre méthode, `M-QuBE`^{3}, conçue pour l'exploration et la visualisation de réseaux multi-couches en sciences humaines et sociales.

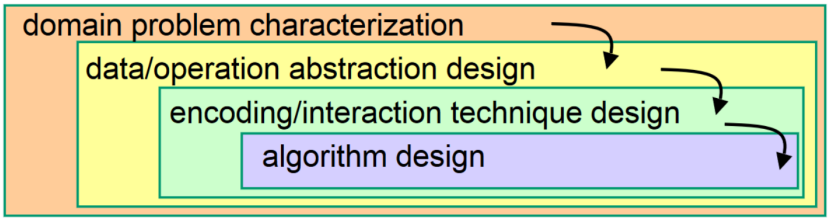

FIGURE *3.6* – ***Nested model*** *: Ce modèle se compose de quatre couches imbriquées décrivant les différentes étapes nécessaires à la conception d'une visualisation performante. L'objectif est de mener une réflexion "en entonnoir" où l'on commence par formaliser le problème puis analyser de plus en plus en détail le contexte et les besoins jusqu'au design final de l'algorithme à utiliser. Image provenant de [61].*

La conception de `M-QuBE`^{3} suit le "nested model" de Munzner (voir Figure 3.6). Ce modèle propose d'expliciter les différentes étapes nécessaires pour produire une visualisation en accord aux besoins de l'utilisateur tout en évitant les erreurs majeures pouvant y survenir (incompréhension des besoins ou du domaine, présentation des mauvaises informations et de la mauvaise manière, algorithme inefficace). La première étape est de prendre connaissance des spécificités des problèmes et des données de l'utilisateur. Cette formalisation s'établit notamment vis à vis des objectifs personnels de l'utilisateur mais aussi du domaine dans lequel il évolue. Cette étape a été réalisée directement avec nos collègues historiens afin d'avoir l'idée la plus précise possible du contexte et des challenges auxquels il est nécessaire de répondre. Dans une seconde et troisième étape, on définit une abstraction

sur les types de données et les opérations qui vont être utilisées pour répondre à ces objectifs avant d'établir la charte graphique, les techniques et les interactions qui vont donner corps aux opérations souhaitées. Ces étapes ont été réalisées en travaillant directement à partir des données (voir chapitre 2) déjà utilisées par le CVCE (https://www.cvce.eu/) et toujours en étroite collaboration avec un expert des données afin de s'assurer de répondre efficacement aux particularités du domaine. Enfin, les algorithmes qui générant la visualisation sont conçus, implémentés et validés.

Ces étapes sont capitales au bon déroulement de toute visualisation : s'appuyant directement sur les données et les contraintes du domaine, elles permettent de conceptualiser et développer la visualisation au plus près des besoins de l'utilisateur. La méthode M-QuBE^3 répond ainsi à l'échelle centrée sur les individus par une approche *bottom-up* et *divide and conquer* où l'aspect itératif et arborescent de la méthode de travail des experts est exploitée par M-QuBE^3 via une boucle d'interaction elle même itérative dont les traces arborescentes sont conservées et exploitables (voir Chapitre 4). Le caractère multi-couche est quant à lui utilisé à travers eScore, une mesure d'intérêt prenant en considération les différentes couches du réseau, permettant ainsi à M-QuBE^3 de pouvoir répondre à l'entièreté des besoins du domaine (voir Chapitre 5). Enfin, M-QuBE^3 et eScore sont évaluées par des experts des données les découvrant pour la première fois afin de juger leur efficacité à exploiter et explorer les réseaux multi-couches en sciences humaines et sociales (voir Chapitre 6).