Algopol, sur les traces du partage

70d5f830 aime que 6ff272cb ait indiqué aimer31769184.

Story dans les métadonnées de Facebook

Introduction au terrain Algopol

Comment on orpaille cette gigantesque masse de merde,
pour en sortir les quelques petites choses qui ont de la valeur ?

Alain Damasio[1]

Facebook semble être dans l'angle mort de la recherche. Cette formule peut paraître provocatrice : elle oublie de nombreux travaux se penchant, d'une manière ou d'une autre, sur les sociabilités numériques. Elle exprime toutefois la difficulté réelle à travailler sur les pratiques elles-mêmes. Les chercheurs utilisent des données issues de questionnaires, où les internautes indiquent leurs usages réfléchis plutôt que leurs pratiques réelles. Des aspirateurs de données donnent matière à des traitements quantitatifs, où l'entrée par les contenus perd les individus. Ce qu'il se passe concrètement sur la plate-forme ne peut être observé qu'indirectement par ces méthodologies. Comme si la recherche tournait autour du dispositif sans pouvoir y rentrer. Certes, ce n'est pas la plate-forme Facebook en elle-même qui présente un intérêt. Mais peut-on se passer d'une description empirique des usages pour prétendre à l'analyse et la critique des pratiques sociales numériques ? Le cas particulier du partage d'information est un exemple criant des limites d'une recherche sans l'étude des usages. Les compteurs informatiques agrègent des clics, les enquêtés disent le sens de leurs clics, mais sans une réunion des compteurs et des enquêtés, les questions tournent en rond. L'enjeu est donc de constituer un corpus pour étudier les pratiques du partage d'information sur Facebook à partir des données d'activité réelle.

Algopol[2] réunit dans un projet ANR des chercheurs en informatique[3] et en sciences humaines, autour de travaux sur « les politiques des algorithmes ». Troisième « saison » d'une série de projets animés par Dominique Cardon et Camille Roth, ce projet prolonge les études précédentes sur la coopération en ligne (Autograph) et la notoriété en ligne (Webfluence). En suivant l'histoire du web, Algopol s'attelle aux réseaux sociaux. C'est dans ce cadre qu'une enquête particulièrement originale a été élaborée : une application qui collecte les comptes Facebook d'un large échantillon d'enquêtés. Ce projet mêle des collaborations humaines, des activités opérationnelles et mes questions de recherches

[1] Auteur de science-fiction, à propos de la vie numérique et de sa mémoire sans fin. Cité par Rue89 et In Limbo, http://rue89.nouvelobs.com/rue89-culture/2015/03/05/conversation-alain-damasio-debarrasser-memoire-258043, consulté le 05/03/2015

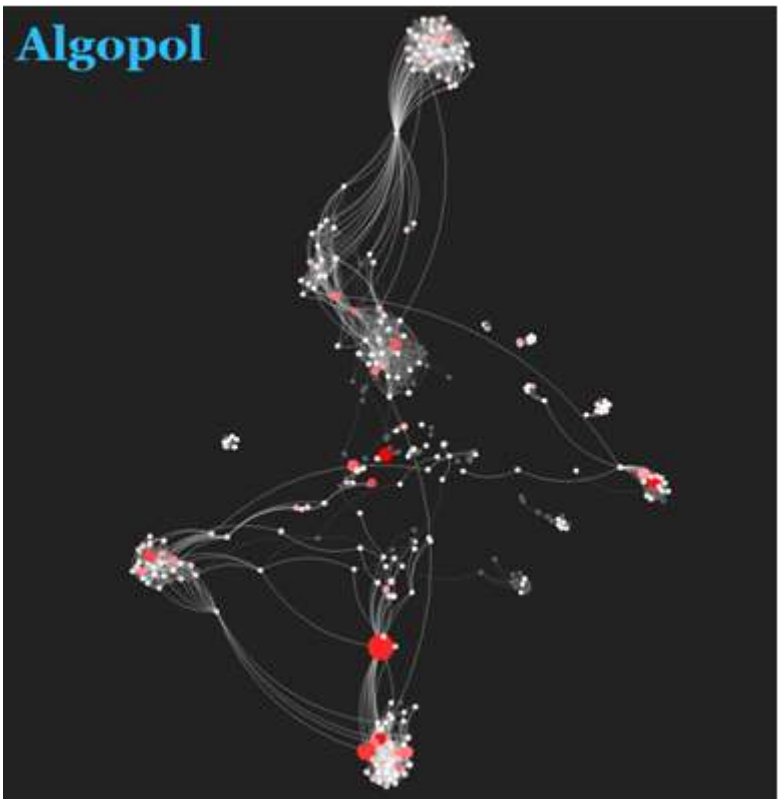
[2] http://algopol.fr, consulté le 30/10/2014

[3] Ma thèse s'est inscrite dans ce projet et a donc bénéficié des nombreux apports théoriques et pratiques des chercheurs du LIAFA, du CAMS, et de Linkfluence

sur le partage d'information. Je vais donc introduire cette partie en « racontant » ce qu'est l'application Algopol et comment s'est déroulée sa mise en œuvre, afin d'apporter une (petite) pierre à l'édifice des multiples projets et recherches qui collectent des traces numériques.

L'application Algopol propose aux enquêtés de visualiser leur réseau d'amis Facebook sous la forme d'une carte interactive. La figure 18 représente la carte d'un enquêté, qui sera appelé Camille. Sur cette carte, chaque point est ami avec Camille sur Facebook ; deux points sont liés et rapprochés si les amis sont aussi amis entre eux. En passant la souris sur un point, le nom de l'ami représenté apparaît. Camille n'est pas sur la carte mais visualise avec Algopol son histoire sociale : le groupe des amis du lycée au milieu, l'équipe de foot du samedi matin à droite, la famille tout en haut, etc. La carte est interactive au sens où l'enquêté peut paramétrer la fenêtre temporelle d'observation de ses amis Facebook, ainsi que la taille et la couleur des points en fonction de critères sur les amis. Par exemple la taille peut être fonction du nombre de *likes* que l'ami a donné aux statuts de Camille et la couleur différente en fonction du sexe de l'enquêté.

Figure 18 : *Carte Algopol de Camille*



De manière générale en sciences, un « réseau » ou « graphe » est constitué de nœuds reliés entre eux par des liens, ou arêtes. L'analyse consiste à observer les graphes en termes de *clusters* (ou groupes de nœuds reliés), ponts (nœuds entre les *clusters*), densité (nombre de liens entre les nœuds au sein du réseau, au sein des *clusters*), et de nombreux autres indicateurs issus de la théorie des graphes (Scott, 1992). Cette formalisation de base s'utilise dans différents contextes : les réseaux du web où les nœuds sont des sites et les liens les citations d'un site par un autre[1], les réseaux académiques où les chercheurs sont des nœuds et les articles auxquels ils ont conjointement collaboré des arêtes, les réseaux économiques où les organisations sont les entités liées soit par des flux financiers soit par des administrateurs communs, et bien d'autres projets utilisent donc les réseaux pour structurer et explorer une problématique[2]. La représentation des sociabilités sous la forme d'un graphe ne date pas de Facebook : en France, Claire Bidart utilise ces représentations pour expliciter avec ses enquêtés l'évolution de leurs groupes de sociabilités (Bidart *et al.* 2011) ; Dominique Cardon et Fabien Granjon ont initié l'utilisation de ces formes comme support d'analyse à des pratiques culturelles (Cardon, Granjon, 2002). Des indicateurs structurels des graphes servent donc à décrire les formes de sociabilités.

Dans la carte Algopol, le réseau représente les groupes de sociabilités à partir des liens déclarés dans Facebook. Stéphane Raux, chercheur en informatique à Paris 7 et Linkfluence, avait connaissance de Sigma.js[3], une API Open Source de construction de graphe. Son idée était d'appliquer cette API avec en entrée les liens d'amitiés sur Facebook. Les nœuds sont donc les amis (alter) de l'enquêté (ego) sur Facebook et les liens sont les amitiés déclarées entre les amis sur la plate-forme. Sigma.js spatialise le graphe en rapprochant les nœuds qui sont liés et en éloignant les nœuds qui ne sont pas liés, formant une carte esthétique et illustrative des groupes d'amis numériques de l'enquêté. Pour construire la carte, l'application Algopol collecte des éléments du compte Facebook de l'enquêté, avec son consentement et en l'informant de l'utilisation qui en sera faite. La visualisation de la carte est donc un levier de recrutement pour faire participer le plus d'utilisateurs possibles au projet et leur demander de confier leurs données aux chercheurs.

Par rapport à une enquête de terrain comme celle menée en lycée, ce type de dispositif pose trois problèmes. Le premier est qu'un certain nombre de partis pris doivent être pris en amont de l'ouverture de l'application, lors de sa conception, et qu'il ne sera pas possible de revenir dessus par la suite. Je présenterai donc dans un premier temps le déroulement de l'application à partir des principes de conception adoptés. Le second problème est qu'une enquête numérique sur les usages numériques met en abyme les

[1] Voir par exemple le projet ediaspora mené par Dana Diminescu, qui représente les cartes web des migrants de différentes origines : http://www.e-diasporas.fr/, consulté le 18/11/2014

[2] Le RT26 de l'AFS anime les recherches sur l'analyse des réseaux sociaux.

[3] http://sigmajs.org/, consulté le 04/11/2014

pratiques, c'est-à-dire que l'utilisation du dispositif dit aussi quelque chose de l'utilisation de Facebook. Collecter les réactions autour de l'application est nécessaire, afin d'identifier les biais de recrutement des participants. La carte permet aussi à l'enquêté de raconter ses activités en ligne et sert à former des hypothèses pour l'analyse des données. Je restituerai donc dans un deuxième temps des éléments captés sur la réception de l'application. Enfin, comme les données collectées sont construites pour et par Facebook et non pas pour répondre aux questions qui animent cette recherche, le travail des données est une étape préliminaire à l'analyse où se posent de nombreuses interrogations (et où les réponses ne se trouvent pas toujours). Je terminerai donc en revenant sur cette préparation qui consiste à faire rentrer des données Facebook dans un Excel, sans perdre ni dénaturer les informations… ce qui peut ressembler à de l'orpaillage.

Trois partis-pris pour construire l'application

L'application[1] Algopol est accessible depuis un ordinateur[2], elle se présente comme une page web et propose un parcours de pages en pages pour participer à l'enquête et visualiser la carte de son réseau. L'architecture fonctionnelle de l'application a du intégrer les contraintes de la recherche, les conditions légales pour travailler sur ce type de données personnelles, et enfin des contraintes ergonomiques et « web » pour que les utilisateurs de Facebook participent. Le parcours de l'enquêté sur l'application est présenté dans l'annexe 7. Trois partis pris ont présidé à la mise en œuvre de l'application et vont être décrits ici, en lien avec les composants fonctionnels mis en œuvre.

Partir des individus, plutôt que des contenus

L'espace numérique a donné un champ d'observation fascinant à de multiples travaux. Les *websciences* s'intéressent par exemple à la construction d'une connaissance dans un espace ouvert de collaboration comme Wikipédia, à la diffusion du bouche à oreille promouvant les œuvres culturelles, ou encore aux développements d'argumentaire dans les controverses grâce aux outils numériques. En prenant l'exemple de la conférence de référence en *websciences*, ICWSM, sur les onze corpus mis à disposition en 2013, sept

[1] Juste pour lever une ambiguïté : le dispositif est appelé « application », car l'enquêté est engagé de manière interactive dans l'exploration de son réseau social. Néanmoins ce n'est pas une « application » au sens d'un applicatif mobile, qui s'installe sur un smartphone.

[2] Le projet n'avait pas le budget pour développer la version mobile de l'application Algopol, ce qui a probablement perdu certains enquêtés qui auraient vu une information sur leur fil Facebook. L'usage lui-même de l'application n'a toutefois pas les caractéristiques d'un usage en mobilité : il faut 5 à 10 minutes continues pour « suivre » le parcours, l'exploration de la carte est plus aisée avec une souris qui permet de zoomer / dézoomer pour voir ses amis, l'exploration est un acte unique puisque la carte ne changera pas sauf si l'enquêté a ajouté / supprimé des amis. Le choix de ne pas développer de version mobile est donc cohérent avec les usages, même si il génère un frein au recrutement.

sont constitués de tweets, un porte sur YouTube et un autre sur Anobii[1]. Tous ces corpus sont construits à partir d'entrées par des objets numériques publics[2] : des tweets, des vidéos, des profils. S'il est donc possible, avec des outils de *crawling* du web, de partir des documents pour explorer l'espace numérique, il est par contre extrêmement difficile de revenir ensuite aux individus produisant cet espace. Par exemple, en observant le web vu des médias dans le chapitre I-2, les commentaires des articles tout comme les performances sur Facebook de ces contenus ne donnaient pas d'information sur les auteurs et *likeurs*, rendant impossible une analyse des profils actifs sur l'information en ligne. Partir des individus, et plus particulièrement des individus ordinaires qui conversent avec leurs amis sur Facebook, inverse donc la perspective adoptée par de nombreux travaux. C'est le premier parti pris de ce chantier Algopol.

Cette entrée complexifie le dispositif, puisqu'elle nécessite de recueillir le consentement de l'enquêté et de se conformer à la législation sur la protection des données personnelles[3]. La législation française impose en effet plusieurs obligations à tout système construisant et traitant une base de données liée à des individus. Il convient de : (1) recueillir le consentement des enquêtés, et leur garantir un accès à leurs données ainsi que la possibilité de se retirer ; (2) sécuriser l'accès à ces données personnelles. La deuxième obligation relève plus de l'architecture technique et ne sera pas abordées ici. Un point sur l'anonymisation des données sera fait dans le paragraphe 3, puisque c'est un des traitements effectué sur la base. Par contre la première obligation sur l'information des enquêtés relève bien du parcours de l'application.

La première étape de l'application Algopol tente d'être la plus explicite possible sur l'objectif et les conditions de la recherche afin de recueillir le consentement de l'utilisateur pour participer à l'enquête. Une fois son accord donné, l'enquêté doit s'authentifier sur Facebook et valider auprès de la plate-forme qu'il donne accès à Algopol à une liste de champs concernant son compte : son profil, ses pages *likées*, ses amis[4], etc. Facebook affiche l'ensemble des champs auxquels Algopol demande accès à l'API, et l'enquêté doit accepter de donner cet accès comme il accepte les conditions générales d'utilisation d'un service, c'est-à-dire tout en bloc. Le consentement de l'utilisateur est donc formalisé deux fois, dans l'application et dans Facebook.

[1] Pour les corpus de données mis à disposition pour la conférence, voir http://icwsm.org/2013/datasets/datasets/, consulté le 18/11/2014. Anobii est un site de critiques de livre.

[2] Un corpus s'intéresse à Facebook mais explore les pages de personnalités politiques ; cette recherche met alors en lumière non pas l'usage générique de la plate-forme mais un usage particulier.

[3] On peut remarquer que le droit français prévoit aujourd'hui des exceptions au droit d'auteur à des fins pédagogiques et de recherche, mais ne prévoit pas d'exceptions au droit sur la protection des données personnelles pour ces mêmes activités.

[4] Cette liste est formalisée par Facebook à travers les requêtes de l'application ; les champs indiqués sont nombreux et en « langage » Facebook.

Le dispositif technique de participation des enquêtés n'est pas problématique en tant que tel. Il créée un biais de recrutement, puisque des internautes renoncent à participer à la recherche en voyant la liste des informations auxquelles l'application va accéder, mais ce biais est attendu. Par contre la participation indirecte des amis a nécessité un dispositif spécifique. Pour construire la carte de l'enquêté, Algopol collecte les noms, le sexe, l'âge, la ville d'habitation des amis ; dans le mur de l'enquêté, l'application a accès aux commentaires ou publications de ces alters ; or ceux-ci n'ont pas explicitement accepté de participer … Pour collecter légalement des données sur les alters, Algopol délègue la mission d'information aux enquêtés. Après son double consentement, l'enquêté voit donc apparaître une page lui demandant de prévenir ses amis de leur participation indirecte. Trois formules sont proposées : (1) publier un message sur son mur ; (2) envoyer un message à certains amis dans leur inbox ; (3) prévenir ses amis par ses propres moyens[1]. Soyons réalistes : aucune de ces solutions ne garantit que chaque ami soit prévenu. Et la troisième solution est bien évidemment hypocrite, dans les échanges avec les enquêtés certains indiquaient explicitement qu'ils ne voulaient pas que leurs amis sachent qu'ils avaient participé. Mais cette solution, élaborée avec la CNIL, permet de réaliser une enquête dans l'enquête sur la responsabilité perçue des internautes vis-à-vis de leurs données et de celles de leurs liens. En effet, l'API Facebook donne accès aux informations sur les amis d'un profil car la réglementation américaine sur la gestion des données personnelles n'oblige pas à ces missions d'informations des utilisateurs. De nombreux services, des jeux comme des médias, peuvent donc utiliser leurs fans pour constituer une base et recruter parmi les amis de leurs fans. En équipant l'enquêté pour prévenir ses amis de leur participation indirecte au projet, l'application Algopol assure donc une forme de sensibilisation des internautes à la gestion de leurs données.

Collecter des données empiriques plutôt que des données déclaratives

A partir du consentement des enquêtés pour que l'application accède à leur compte, Algopol utilise l'API de Facebook, c'est-à-dire que l'application dialogue avec Facebook pour récupérer les informations du compte utilisateur. Algopol collecte très exactement trois types de données : le profil de l'enquêté[2], son réseau social c.à.d. le profil de ses amis, et son mur. Le deuxième parti pris d'Algopol est donc que de collecter des données d'usage empiriques, et non pas des données déclaratives.

Deux structures sont les principales pourvoyeuses d'études sur Facebook aujourd'hui : le Pew Internet Research Center et l'équipe de Charles Steinfield et Nicole Ellison à

[1] A la date de rédaction de ce manuscrit, les analyses de l'utilisation de ce dispositif ne sont pas disponibles.

[2] Sauf trois champs soumis à une réglementation plus stricte sur la sécurisation des données : les croyances religieuses, l'opinion politique, et l'orientation sexuelle. Si Algopol avait collecté ces trois champs des profils Facebook, l'application aurait dû être déclarée à la CNIL dans une catégorie plus élevée. Sans data sur ces champs, il n'est pas possible de savoir quelle part d'enquêtés les renseigne, et encore moins quelle part d'enquêtés les renseigne de manière ludique.

l'Université du Michigan[1]. Toutes deux ont initialement travaillé avec des questionnaires, administrés à un panel d'internautes américains ou aux étudiants de Chicago, et prolongés par des entretiens ou des questionnaires approfondis. Il n'est bien sûr pas question de mettre en doute ici l'intérêt et la qualité de ces travaux, qui contribuent de façon déterminante à une meilleure connaissance des usages, des pratiques, et des impacts sociétaux des activités numériques. Mais toutes deux testent aussi aujourd'hui des manières d'accéder aux usages réels : dans l'enquête de 2012, le Pew Internet Research a ainsi eu l'accord de 269 enquêtés, sur 877, pour récupérer leurs données auprès de Facebook (Hampton *et al.* 2012) ; Nicole Ellison et son équipe ont travaillé avec Facebook pour analyser le contenu des messages publics postés sur la plate-forme (Ellison *et al.* 2013). Et enfin, la *data team* de Facebook monte en puissance et produit elle-même des recherches[2]. Analyser les usages réels à partir des données des utilisateurs constitue ainsi un champ en cours de construction de la recherche.

Un dispositif exploratoire avait été mené chez Orange Labs en 2010. Il s'appuyait sur un profil d'enquêteur virtuel, appelé Julie Tagline. En devenant ami avec Julie Tagline, les enquêtés acceptaient que celle-ci collecte les informations qu'ils publient en *crawlant* régulièrement leur *timeline*. Cette méthodologie avait montré l'intérêt de partir des données mais aussi des contraintes techniques très fortes : la *timeline* de Facebook évoluant très régulièrement, la maintenance du *crawler* nécessitait une surveillance permanente du développeur ; la personnalisation des contenus affichés sur Facebook n'assurait pas au robot la collecte systématique des publications des enquêtés. Algopol a résolu ce problème en collectant non pas l'activité de l'enquêté *à venir*, à partir de l'accord de l'enquêté, mais l'activité *antérieure* sur le mur. L'API donne ainsi les données d'usage réel, et presque « officiel » : il n'est bien entendu pas possible d'avoir accès aux messages effacés[3], comme il n'est plus possible d'avoir accès aux amis reniés. Ce sont donc des données passées de l'histoire de l'enquêté que l'application Algopol récupère pour photographier les usages de Facebook.

Ne pas essentialiser la vie en ligne

Enfin, après avoir collecté les données du compte de l'enquêté et avant d'afficher la carte, Algopol pose quand même quelques questions à l'utilisateur... Ces questions ne sont pas factuelles sur ses activités en ligne, mais s'intéressent à des éléments hors ligne que les data ne peuvent pas dire. Ainsi, un formulaire « ego » demande à l'enquêté son sexe, son âge, son code postal et sa profession, afin de pouvoir intégrer les caractéristiques sociodémographiques des enquêtés dans la description des profils. Puis un formulaire va

[1] Pour ne citer que quelques études, voir Boase *et al.* 2006, Hampton *et al.*, 2011, Steinfield *et al.*, 2008, Ellison *et al.*, 2008, Ellison *et al.*, 2012.

[2] Voir http://www.facebook.com/datateam.

[3] Une recherche de Das & Kramer s'intéresse tout de même à l'autocensure sur Facebook, mais grâce à des données obtenues directement auprès de Facebook et non pas par l'API (Das, Kramer, 2013).

poser cinq questions sur le lien de l'enquêté avec ses cinq « meilleurs » amis Facebook. Le troisième parti pris de l'application Algopol est donc de ne pas essentialiser la vie en ligne en oubliant les déterminants sociaux et les pratiques hors lignes, mais bien de lier les pratiques des deux univers.

Les recherches de Facebook, publiées par exemple par Jones *et al.* montrent que les « meilleurs » amis sur Facebook sont aussi les « meilleurs » amis hors ligne (2013). A partir de 789 dyades décrivant un enquêté et un de ses meilleurs amis, les auteurs ont pu analyser les activités réciproques de ces liens. Ils montrent que les interactions publiques comme *liker* et *commenter* le statut d'un ami sont des indicateurs satisfaisants de la force du lien entre les interactants, et que les indicateurs liés aux activités privées comme les messages échangés en inbox ne sont pas plus utiles pour inférer la force du lien. L'application Algopol s'appuie sur ce résultat pour qualifier l'usage de Facebook de l'enquêté. Elle présente à l'enquêté les deux amis qui ont le plus *liké* ses statuts, les deux amis qui ont le plus commenté ses statuts, et l'ami avec qui il a le plus d'amis en commun[1]. Bien sûr si ces amis étaient redondants, les performeurs suivants dans le hit-parade étaient présentés, pour que l'enquêté décrive bien cinq profils. Pour chaque ami, l'enquêté devait indiquer des informations sur l'ancienneté de la relation, la fréquence des contacts, l'affection ressentie pour la personne sur une échelle de 1 à 5 et la nature de la relation (ami, connaissance, famille, collègue, autre). En admettant les résultats de Jones, si l'enquêté a pour plus proches amis sur Facebook des collègues pour qui il n'éprouve pas spécialement d'affection, alors cet enquêté a un usage de Facebook professionnel.

A l'issue des formulaires sur les cinq amis, l'enquêté peut (enfin) accéder à la carte de son réseau social. Nous allons voir maintenant les réactions qui découlent de cette visualisation.

Le recrutement, l'expérience utilisateur, le self-data

Une fois que l'application est en place, appuyer sur le bouton « *on* » parait somme toute assez simple : il suffit de supprimer le blocage par mot de passe qui empêchait l'accès à l'application pendant les tests. Mais encore faut-il trouver des participants … Trois filières de recrutement ont été utilisées. La première a été activée en amont de l'ouverture grand public : l'institut de sondage CSA a proposé à ses panélistes internautes de faire l'application, à l'issue d'une enquête sur les usages de la *social TV*. Cette filière a constitué un échantillon de 877 enquêtés, représentatif des utilisateurs de Facebook en France. L'application a ensuite été ouverte au grand public grâce à un article du Monde.fr

[1] Sauf que la sélection n'a pas toujours porté sur l'ensemble des amis et l'ensemble de l'ancienneté du compte, notamment aux périodes de forte affluence pendant lesquelles l'application ne collectait que 100 amis avant d'afficher la carte.

publié le 12 décembre 2013 intitulé « Quand la recherche like Facebook »[1] : cet article inventorie les applications de recherche qui utilisent Facebook, dont Algopol puisque l'équipe projet avait été interviewée par le journaliste David Larousserie. Cette communication a généré un très fort trafic sur l'application, puisque le site a reçu en deux jours pas loin de 10.000 visites[2]. Enfin, l'équipe du projet a lancé des appels à participation auprès des réseaux de chercheurs, ce qui nous a permis de recruter probablement nos collègues amis, quelques amis d'amis, et surtout un très grand nombre d'étudiants ! En effet, certains enseignants-chercheurs utilisaient Algopol dans leur cours sur l'analyse des réseaux sociaux, en illustration de séminaires sur les sociabilités numérique, ou encore en exemple de *websciences*. Au 17/11/2014, l'application comptait plus de 12.000 participants, dont 75 % d'hommes et 4000 étudiants. Une ultime vague de recrutement a pu être réalisée fin avril 2015. Une évolution de l'API Facebook prévue le 30/04/2015 a effectivement rendue l'application inopérante, puisqu'il n'est plus possible d'accéder aux liens entre les amis d'un enquêté. Nous avons sorti les premières analyses du corpus pour susciter la participation de profils sceptiques sur l'utilisation de leur compte, en tentant notamment de cibler les femmes. David Larousserie a ainsi présenté dans un article « Les trois grands profils d'usage de Facebook »[3] que nous avions identifiés dans les analyses préliminaires[4]. Cet article a généré 3600 participations supplémentaires. Au total, le projet Algopol a donc collecté 16439 comptes Facebook.

En parallèle de cette diffusion de l'application, nous avons mis en place un certain nombre de dispositifs pour échanger avec les enquêtés, autant dans une logique de service que dans une logique ethnographique. Le but était d'accompagner le recrutement, de résoudre les problèmes remontés par les enquêtés, et d'en « profiter » pour capter des retours sur l'expérience utilisateur. Nous avions bien sûr une adresse mail sur laquelle les enquêtés pouvaient nous contacter ; elle a reçu principalement des demandes de

[1] http://www.lemonde.fr/sciences/article/2013/12/12/quand-la-recherche-like-facebook_4332566_1650684.html, consulté le 12/11/2014

[2] Ce pic de charge a forcément bousculé l'application … Pour les enquêtés qui se sont connectés le 12 et le 13 décembre 2013, l'application n'a pas pu afficher leur carte car Facebook n'allouait pas assez de « jetons » à l'application pour que nous puissions collecter les informations des comptes. Nous avons mis deux semaines à rattraper le retard pris pour les *early adopters* de ces premiers jours, sachant que Stéphane Raux a très rapidement mis en place une solution technique pour que les nouveaux visiteurs soient traités au plus vite plutôt que d'allonger la liste d'attente, et que nous avons abondamment communiqué (sur Facebook, sur Twitter, par mail) pour nous excuser du retard pris et pour indiquer la procédure pour revenir ultérieurement. A la réflexion, je reste étonnée du peu de critiques que nous avons eues… au contraire les enquêtés ont semblé particulièrement compréhensifs et même encourageants. Une mention spéciale à un « ancien », sous pseudonyme sur Facebook, qui nous a consacré quelques heures pour nous aider à fixer un bug lié à la version 3.2.8 (ou autre !) du navigateur Internet Explorer.

[3] http://www.lemonde.fr/pixels/article/2015/04/17/une-etude-revele-les-trois-grands-profils-d-utilisateurs-sur-facebook_4618227_4408996.html, consulté le 11/05/2015.

[4] http://algopol.huma-num.fr/appresults/premiers-resultats/, consulté le 11/05/2015.

retraits mais aussi des questions[1]. Nous animions une page Facebook, pour diffuser à chaud les nouvelles de l'application au lancement et à froid des résultats[2] ; cette page a aussi reçu des messages, public et privé. Enfin, nous avons pu participer à différents évènements liés au numérique et à la recherche : Futur en Seine, le festival du numérique, qui s'est tenu du 12 au 16 juin 2014 à Paris ; la fête de la Science, à l'université Paris Diderot, où nous étions présents du 8 au 11 octobre 2014. Ces échanges et présentations donnent des éléments sur l'expérience que les enquêtés ont d'Algopol. Or les pratiques de l'application indiquent en creux les pratiques de Facebook et montrent les ombres des usages que nous ne captons pas avec les données. Sans prétendre faire une réelle enquête de réception, voilà quelques éléments des enquêtés sur leur participation à l'application.

Donner son corps (Facebook) à la recherche ?

L'intérêt des présentations est que ces situations de recrutement permettent d'entendre les refus des personnes qui ne souhaitent pas participer. Le premier argument entendu est « *si vous collectez mes données, pas question* ». Cette réaction est justifiée avec un mélange plus ou moins confus de peur de la surveillance et de sensibilité très forte au caractère personnel des données. Qu'Algopol fasse craindre une surveillance est inhérent au fait de partir des individus et non pas des contenus. Cette question ne se pose pas quand une recherche collecte les flux RSS des médias ou les pages de conversation sur Wikipédia. Il est vrai que l'application collecte de nombreuses informations sur un enquêté, comme son corps virtuel. La recherche n'utilise pas l'information personnelle d'un enquêté, mais les informations restituées parmi toutes celles d'un corpus d'enquêtés. De même qu'un patient se déshabille devant un médecin pour lui montrer non pas sa nudité mais ses symptômes, Algopol demande aux utilisateurs de Facebook de révéler leurs comptes non pas pour les exposer mais pour dessiner les formes idéal-typiques des sociabilités numériques. Le travail de design de Rose Dumesny, pour la participation d'Algopol à Futur en Seine, explicite cette recherche de formes et de régularités de manière ludique avec des carottes et des artichauts (voir annexe 8). Ces arguments ont parfois permis de donner confiance à des visiteurs qui n'auraient pas participé de prime abord, mais il faut reconnaître que convaincre les profils frileux n'a pas toujours fonctionné. L'échantillon souffre donc très probablement d'une sous-représentation des comptes Facebook intimistes.

Les internautes sont-ils inquiets sur l'utilisation de leurs données personnelles ? Le projet n'adresse pas directement cette question, mais quelques éléments de contexte permettent de resituer l'environnement ambiant de l'application. Tout d'abord, les révélations d'Edward Snowden ont mis en lumière des pratiques de surveillance à grande

[1] Au 04/11/2014, le mail contact@app.algopol.fr a reçu 76 mails

[2] Au 04/11/2014, la page Facebook de l'application a 785 fans (et l'application a été *likée* par 3323 internautes).

échelle et non officielles. A une échelle plus petite, le web grouille de transferts de données entre services[1], notamment pour monétiser les applications gratuites avec de la publicité ciblée. Il est donc légitime de s'inquiéter de pratiques non encadrées par une législation et non consentie par les individus. Cependant, l'inefficience de ces pratiques semble montrer que le traitement des données n'est pas (encore) aboutit. Ce fourvoiement résulte de deux constats actuels.

D'abord, les activités numériques ne disent qu'une partie des pratiques d'un individu. Les données incarnent une manne pour l'observation des usages numériques mais un leurre pour l'observation des individus. L'incertitude de la valeur des données personnelles sur Facebook se retrouve dans la confrontation de deux commentaires, postés par des internautes sur la page Facebook du monde.fr suite au statut mentionnant le lancement d'Algopol[2].

Figure 19 : *Données personnelles partielles ou entières*

Dans le premier commentaire, une femme (d'après son identifiant Facebook) signale que les internautes ont des stratégies de publication sur les réseaux sociaux, semblant sous-entendre que toute analyse des comportements numériques est vaine car amputée de certaines pratiques retenues ou cachées. Dans le second commentaire, une autre femme souligne le fait que l'application accède à « toutes » les données, et qu'une grande entreprise participe à cette recherche, ce qui range Algopol dans la même famille que les entreprises qui participent à des formes de surveillance secrète et massive des activités en ligne. Pour l'une, les données collectées sont partielles et ne disent rien de valable sur les individus ; pour l'autre, les données exhaustives en disent trop … Il paraît simpliste de

[1] Cookievizz, développé par la CNIL, permet de visualiser au cours d'une navigation quels sont les sites avec lesquels le navigateur échange des informations. Voir différents outils de ce type dans l'article suivant : http://rue89.nouvelobs.com/2014/08/30/grace-a-donnees-peut-tout-savoir-voyez-meme-254336, consulté le 18/11/2014

[2] A noter que ces commentaires étant postées sur une page publique, j'aurai pu ne pas les anonymiser …

dire que la vérité est entre les deux. Et pourtant, c'est l'expérience que j'ai ressenti en explorant les données : l'impression d'être indiscrète et de « voir » les gens, et en même temps la perception de ne rien pouvoir comprendre sans les voir vraiment, ou sans les voir parmi tous les enquêtés. Ce n'est pas parce qu'une API permet de collecter tout un compte que les systèmes deviennent pour autant panoptiques. L'action, notamment en ligne, ne dit pas tout de l'acteur.

C'est le deuxième constat sur l'échec de la surveillance qu'on observe alors. Croire que les données Facebook, une fois accédées, donnent une vision limpide et claire de chaque participant est une gageure, pour preuve j'aurai dû compter les heures de travail passées à tenter de comprendre quelque chose dans la masse de données du projet. Le fantasme de l'analyse des traces numériques donne encore plus de pouvoir aux machines que ce qu'elles n'ont aujourd'hui réellement. Ainsi, le projet Algopol peut témoigner que la vie privée des enquêtés n'est pas (encore) mise à nue dans leurs données Facebook. Comme le montre les travaux d'Antonio Casilli, l'attention à la protection de la vie privée n'a jamais été aussi forte, au point de la rendre bien vivace (Casilli, 2013).

En testant cette question de la surveillance auprès des participants à Algopol à Futur en Seine, certains volontaires ne voyaient pas le « problème ». Ces personnes expliquaient qu'elles exerçaient un contrôle en amont sur leurs données : ce qui est mis sur Facebook, expression ou photo, est à leurs yeux une expression publique et donc préalablement filtrée. Entre les deux extrêmes de « mes données sont privées » et « mes données sont publiques », il y a tout un continuum de rapports au caractère sensible ou non de ce qui se passe sur Facebook. Et cet éventail touche les différentes activités possibles sur Facebook. Une jeune femme a par exemple réagi à la liste des informations collectées par l'application en s'interrogeant sur la collecte des photos : « *Vous allez prendre toutes mes données, même les photos ? C'est chaud ça…* » (Femme, 23 ans). Par la suite la jeune fille m'indiquera que ce sont les photos avec ses amis qui lui paraissent intimes[1]. Au cours d'un entretien plus complet avec une femme de 33 ans, nous avons exploré son mur Facebook et ses amis, avant d'arriver sur la page listant les pages dont elle est « fan » (qu'elle a *liké*) et l'enquêtée de s'étonner elle-même en concluant : « *finalement, cette page [des pages likées] correspond beaucoup plus à ce que je suis personnellement que la carte de mes amis* ». Le curseur de « sensibilité » d'une information personnelle varie donc d'une information à l'autre pour une même personne, et d'une personne à l'autre pour une même information.

[1] Au risque de rouvrir le sujet de la surveillance, c'est pourtant des données que nous ne pouvons absolument pas analyser ! L'API nous indique que l'enquêté a publié un statut avec une photo et tagué des personnes, mais la photo elle-même n'est pas disponible ; et quand bien même elle le serait, il faudrait des outils d'analyse d'images aujourd'hui encore bien moins avancés que les outils d'analyse textuelle pour pouvoir tenir un propos consistant sur la masse de .jpg constituée.

A partir de ces observations des résistances liées à la collecte des données Facebook par l'application Algopol, il est clair que les enquêtés, en particulier ceux recrutés par bouche à oreille, sont des profils décomplexés sur Internet. Les 16000 enquêtés ont très certainement un biais de recrutement surreprésentant les internautes qui utilisent Facebook avec une posture d'expression publique plutôt qu'avec des pratiques intimistes.

Les interactions sont publiques mais le réseau est privé

Les réactions sur la carte sont aussi une source de *verbatim* intéressants pour comprendre le rapport à Facebook de nos enquêtés. Les enquêtés commencent par identifier chaque groupe, ou les points « en gros », et se mettent à raconter tel ami ou tel lien et tel groupe. Il n'y a rien d'étonnant à ce que les amis de l'équipe de foot soit tous amis sur Facebook, mais le voir permet de percevoir ce groupe parmi les autres amis. Après cette première approche, qui peut prendre du temps, les enquêtés se mettent à utiliser les paramètres comme la période d'observation pour se souvenir de qui étaient les premiers amis, ou la taille des boutons pour tester qui *like* le plus les statuts. Cette exploration donne rarement des surprises, j'ai une fois vu un enquêté découvrir un lien d'amitiés entre des alters sans que ce lien ne soit connu au préalable. Au contraire, les enquêtés disent en général percevoir eux-mêmes les « résultats » de l'application : « *je le savais, ma belle-mère me like plus que ma mère !* » (Femme ~35 ans, cadre). L'utilité de la carte peut alors être mise en question, comme l'interroge une amie d'enquêté :

Figure 20 : *« Qu'est-ce que ces algo's t'ont réellement appris(…) ? »*

L'ambiguïté vient du fait que la carte restitue à l'enquêté ses amis, qu'il connait et fréquente sur Facebook ; donc la carte est sensée être sue, à tout le moins perçue.

Si la carte ne produit en général pas de surprise, elle fait tout de même son petit effet. Les amis ne sont en effet pas classés par ordre alphabétique, rangés par fréquence de contact, ou archivés par ancienneté. La carte structure les amis en fonction de leur groupe d'appartenance et donc de la densité des liens. Autant les chercheurs ne peuvent comprendre que la structure du réseau avec des indicateurs, autant l'enquêté lit cette

visualisation avec des émotions, retrouvant des amis appréciés, des connaissances délaissées, des ex reniés mais non effacés, des histoires tristes comme des connaissances décédées. Dans les questions que les enquêtés ont posées à l'équipe projet par mail ou via Facebook, la plupart portait sur le fait que tous les amis n'apparaissaient pas sur la carte[1], quand bien même il en manquait 2 alors que plus de 500 étaient présents. Lors du pic de charge d'avril 2015, un enquêté a voulu exercer son droit de retrait car sa femme n'apparaissait pas sur la carte (à cause du délai de chargement des données) et cela semblait créer une scène de ménage … En testant l'application avec mon propre compte Facebook, j'ai un jour vu s'afficher dans ma carte Algopol des noms de personnes qui n'étaient pas mes amis. J'ai trouvé ce bug étonnamment effrayant : je croise tous les jours des inconnus dans le métro, mais les voir qualifier comme des amis m'a paru très intrusif ! Une enquêtée décrit Algopol comme « flippante » car elle a découvert des « meilleurs amis insoupçonnés » (dans le hit-parade des meilleurs *likers a priori*).

Figure 21 : *« j'ai des « meilleurs amis du monde » insoupçonnés… »*

Ce ne sont pas les meilleurs amis attendus, il manque des amis, ce ne sont pas du tout des amis, ce ne sont plus des amis… Cette analyse des rejets de la carte montre que les participants donnent à Algopol leur profil et leurs expressions, mais reçoivent en retour autre chose. Les réactions plus positives qui consistent à « reconnaître » son réseau et identifier les différents groupes montrent que la carte est un « miroir » des sociabilités. L'application Algopol illustre une formalisation structurelle de la sociabilité avec une perspective jamais vue, celle des groupes d'« amis d'amis ». La carte ne dit pas tant ce que Facebook sait de l'enquêté et que l'enquêté ne sait pas (*sic*), mais ce que l'enquêté perçoit des différences relationnelles et structurelles entre ses amis Facebook, et que les amis de l'enquêtés ne savent pas.

Cette visualisation est ainsi ressentie comme très personnelle, ce qui explique l'échec des « partages » de la carte. Cette fonctionnalité avait été développée afin de générer un recrutement viral sur l'application : l'idée était que si un enquêté publie sa carte, ses amis auront envie de voir la leur et participeront à leur tour. Force est de constater que la

[1] L'application ne peut pas collecter les informations des amis qui ont paramétré leur compte en interdisant à l'API d'ouvrir leurs données aux applications utilisées par leurs amis. Ce paramètre est activé par des utilisateurs probablement experts de Facebook.

viralité de l'application a échoué[1], les enquêtés sont venus des médias mais pas des amis. Et effectivement publier la carte a été largement rejeté par nombre d'enquêtés rencontrés, et tout particulièrement par les femmes. Soit la structure sociale de la carte et l'affection de l'enquêté pour ses amis sont cohérentes, alors Algopol affiche une position relative des amis les uns par rapport aux autres qui se lit comme des choix affectifs. Publier sa carte revient alors à afficher ses préférences affectives, ce qui n'est pas très fin. Soit la structure et l'affection sont distinctes (« *mon meilleur liker, c'est un boulet* », homme 38 ans, cadre), alors la carte Algopol montre une stratégie relationnelle sur Facebook qu'il n'est pas non plus nécessaire de montrer. Ainsi, même pour les enquêtés qui considèrent leurs interactions sur Facebook comme publiques, la structure du réseau est, elle, perçue comme privée.

L'expérience des self-data : règles et ajustements numériques

Ces dires des enquêtés mettent en avant deux traits caractéristiques des sociabilités sur Facebook. Le premier élément qu'il convient de rappeler est que les dispositifs numériques, et notamment Facebook, génèrent un tropisme vers l'activité en ligne. C'est un enquêté qui a souligné ce point en réagissant à la sélection des cinq amis à décrire. Il a hésité un temps avant de comprendre « *en fait, euh, c'est mon Facebook d'avant (…) de quand j'étais actif, et qu'on découvrait le truc et tout* » (Homme, 36 ans, Cadre). Il raconte en effet que ses meilleurs *likers* et *commenters* sont des collègues d'il y a plusieurs années, qu'il voyait quotidiennement au moment où tous se mettaient sur Facebook avec une certaine émulation. Par la suite, l'enthousiasme s'est tassé, l'enquêté a moins publié, ces personnes se sont éloignées, et comme l'enquêté publie beaucoup moins de statuts, les nouveaux collègues, tout aussi sympathiques que les anciens, n'ont pas l'occasion de produire beaucoup de *likes* et *comments*. Ainsi, les amis de la période active restent sur le podium des meilleurs amis Facebook car même s'ils n'augmentent plus leur score, ils ne sont rattrapés que très lentement par les nouveaux amis actifs sur une période plus contemporaine.

Le deuxième élément qui me parait important est que les *data* d'activités personnelles sur Facebook mélange les règles et les dérogations aux règles de l'enquêté. C'est ce qui ressort de plusieurs entretiens sur la réception de la carte et ce qui semble être le levier principal de la réflexivité sur la pratique de Facebook. Les enquêtés commencent en effet par raconter comment ils utilisent Facebook en indiquant les règles qu'ils se donnent, par exemple « ne pas être ami avec le boulot, parce qu'il faut bien séparer les deux, la vie perso et la vie professionnelle » ou « ne jamais parler politique, on sait jamais, ça peut

[1] L'équipe projet a imaginé nombre de « widget de partage », avec parfois un score de sociabilité, parfois un podium des meilleurs amis, parfois des pins de performance. La formule retenue est donc un widget présentant la carte, et le hit-parade des 5 meilleurs amis suivant le paramètre de taille des points et la période d'observation implémenté sur la page au moment où l'enquêté partage la carte ; j'ai ainsi vu quelqu'un partagé 1 carte par année de présence sur Facebook.

fâcher ». Ces règles m'ont souvent paru très formelles et pas seulement chez les enquêtés de milieux intellectuels ; les lycéens de la partie II verbalisent sans hésitations les conditions qu'ils se sont donnés pour devenir ami avec un inconnu par exemple. Cette autorégulation de la pratique peut résulter du discours médiatique et ce qu'on entend des pratiques numériques, ainsi que d'une forme de pédagogie non coordonnée émanant des institutions, des professeurs pour les lycéens mais aussi des collègues, des amis, des expériences pour les adultes. Dire les règles que l'on s'est fixé sur Facebook c'est dire que l'on a entendu et assimilé le bruit ambiant sur les bonnes manières d'utiliser le réseau socionumérique.

L'expérience de *self-data* proposée par l'application Algopol montre aux enquêtés les dérogations aux règles auto-prescrites : un ami qui est isolé parce que c'est un collègue, un inconnu qu'on a accepté pour un flirt, etc. Une enquêtée raconte plus précisément une expérience avec la publication de photos. Elle signale qu'elle n'aime pas mettre des photos de personnes, à part ses photos de profils ; elle insiste deux fois au cours de notre échange sur sa règle de ne pas poster des photos de groupe. Elle prend pour anti-modèle son cousin qui poste les photos de fête de famille, photos qui la mettent mal à l'aise et donc elle établit sa propre règle d'usage à l'inverse. Plus tard dans la discussion, en explorant son historique Facebook, nous passons sur l'onglet Photo et Album, et elle y retrouve un ou deux albums de photos de soirées, où il y a donc des photos de groupes de même nature que celles des fêtes de famille. Elle se justifie un peu en disant que c'était au début de Facebook, qu'elle n'avait pas encore très bien fixé les règles … mais constate avec une certaine humilité qu'elle fait des entorses à son propre règlement. Ainsi, ce que les *self-data* peuvent montrer des pratiques individuelles sert à formaliser les règles auto-fixées et les ajustements tolérés. Algopol permet une réflexivité sur les pratiques personnelles qui supplée à l'absence de formation et de discussion sur les usages.

Faire parler les données

Que les enquêtés puissent raconter leur réseau et lui donner un sens personnel avec leur histoire sociale, soit. Mais comment appréhender les 16000 réseaux ? Et les dizaines de millions de lignes correspondant au moindre clic d'un enquêté ? Une fois que l'application Algopol est développée, ouverte, diffusée, ce n'est que le début… Si le téraoctet[1] de données Algopol fait passer le projet dans la catégorie des *very big data* en sciences humaines et sociales, il convient d'abord de relativiser le matériel de recherche que sont les « données ». Des sociologues se sont penchés sur ces objets pour montrer, dans différents domaines, que les données ne sont jamais brutes mais toujours produites. Le

[1] Stéphane Raux m'a dit que les données Algopol ne faisaient pas 1 To, mais je n'ai pas le chiffre exact… et en tout cas cette exagération rend bien compte du fait que les données sont surabondantes.

travail des données a ainsi été observé par Éric Dagiral et Ashveen Peerbaye *via* la structure et le renseignement des bases de données concernant les maladies rares. Ces chercheurs montrent que les acteurs, aussi bien informaticiens qu'éditeurs et rédacteurs qui saisissent les données, entreprennent des ajustements techniques et pratiques qui deviennent constitutifs de la connaissance obtenue (Dagiral, Peerbaye, 2012). Sur les données de l'administration, Samuel Goëta et Jérôme Denis étudient le mouvement de l'*open data* et le travail des acteurs permettant de rendre ces données de l'administration accessibles à tous ou du moins à certains, par exemple aux développeurs, aux entreprises, ou aux citoyens (Denis, Goëta, 2013). La production de données Algopol pour la recherche, à partir des données Facebook, est donc nécessairement constitutive des questions et analyses qui vont être menées. Il faut donc décrire cette étape préliminaire, quand bien même elle paraît rebutante. Concrètement, comment passer du « json[1] » que les informaticiens ouvrent avec une commande « rs tar z »[2], à un fichier « excel »[3] sur lequel faire des tableaux croisés dynamiques ? Et c'est sans parler de l'analyse textuelle des contenus des messages, qui ne rentrent de toute façon pas dans un Excel. La collaboration informaticien – sociologue nécessite de faire des allers-retours incessants entre le traitement en masse de la base et les données brutes individuelles pour vérifier que les classements et catégorisations créés sont cohérents. Les prochains paragraphes inventorient donc les étapes de constructions des données à analyser.

Figure 22 : *« Données » Algopol affichées sur un écran d'informaticien*

[1] Json est le format de fichier des données Algopol. Un enquêté est décrit avec 4 fichiers json : le fichier « ego » des informations répondues dans l'application ; le fichier « réseau » décrivant le réseau d'ami, le fichier « profil » avec les informations de son compte, et le fichier « statuses » avec le mur.

[2] Commande exécutée par les informaticiens dans une « console » pour ouvrir un fichier, plutôt que « double cliquer » dessus comme pour un fichier Word.

[3] Ou un autre logiciel, je ne promeus pas spécialement la suite Office de Microsoft … mais le lecteur constatera que xls et json ne sont pas du même monde, et pourtant informaticiens et sociologues arrivons à nous entendre.

A la recherche des indicateurs

Si toute recherche se structure à partir d'hypothèses, le travail de terrain impose d'explorer de nombreux indicateurs sans *a priori* sur leur utilisation. Cette indétermination est d'autant plus forte en ligne que les indicateurs sont à construire. Pour analyser les pratiques médiatiques hors ligne, les caractéristiques sociodémographiques comme le niveau de diplôme des enquêtés sont indispensables, puisque le lien entre diplômes et informations a été établi par des précédentes recherches. Mais dans le monde du numérique, il n'existe pas encore de repère comme le diplôme : quels sont les indicateurs qui permettent de décrire le niveau d'un enquêté numérique ? Et surtout, comment concentrer des informations foisonnantes et hétéroclites dans quelques éléments descriptifs ?

Sur les profils, il y a de nombreux champs qu'un utilisateur de Facebook utilise pour se décrire. Le renseignement de ces champs propose soit du texte libre (pour les centres d'intérêt, ou les goûts musicaux par exemple), soit des catégories multiples (pour le sexe, la ville d'habitation)[1]. Autant le texte que les catégories sont difficiles à utiliser car ils sont très hétérogènes et probablement folkloriques. Nous n'avons pas pu collecter les opinions politiques renseignés par les enquêtés dans leur profil, mais rien ne dit que ce champ ait été renseigné sérieusement et donc de manière significative pour des analyses. Sur les 529 enquêtés CSA ayant renseigné les informations du formulaire « ego », 14 (2,6 % de l'échantillon CSA) ont indiqué un sexe différent de celui correspondant à leur profil Facebook. Les champs Facebook de description du profil semblent donc inutilisables … L'indicateur pertinent semble alors être le nombre de champs renseignés : il est significatif de l'investissement que l'enquêté met dans son profil et de l'exposition qu'il accepte. Avoir un profil sans information de sexe, d'âge, de ville, ni statut marital, sportifs préférés, ou films adorés témoigne d'une exposition limitée sur les réseaux socionumériques. Alors qu'un profil avec de nombreux éléments renseignés, que ces éléments soient vrais, faux ou même un peu des deux, témoigne d'un certain engagement. La description d'un profil enquêté passe donc par le nombre d'informations du profil renseignés plus que par les informations elles-mêmes. Ce raisonnement fait sur les éléments du profil d'un utilisateur de Facebook peut être repris sur toute activité numérique : l'entrée qui indique la présence ou l'absence d'un champ est nécessaire, avant de passer à l'évaluation du champ.

Sur les indicateurs liés au réseau social, il y a alors un graphe à étudier et donc une autre dimension... La représentation de la carte Algopol part d'une matrice des liens d'amitié

[1] S'il faut d'ailleurs rappeler la dimension politique des catégories et des données, prenons l'exemple du sexe : Mark Zuckerberg a décidé de soutenir la cause LGBT en proposant plus de 50 désignations pour renseigner son identité sexuelle. Le statut marital propose lui plus de 10 situations, allant du classique « marié » au « partenariat domestique » en passant par l'indéterminé « c'est compliqué ». Et nous n'avons abordé que deux champs, sur la vingtaine possible dans un profil …

entre les alters. Une dizaine d'indicateurs sont utilisés « classiquement » pour mesurer le graphe de chaque individu (par exemple le nombre d'amis isolés, le nombre de communautés détectées par la méthode de Louvain, le diamètre du réseau, le nombre d'amis dans la plus grande composante connexe, etc.). D'autres types d'indicateurs seront envisagés dans le projet, par exemple des indicateurs plus structurels basés sur l'identification de motifs dans un graphe, et la description d'un nœud du graphe en fonction des motifs dans lequel il est impliqué. Mais reconnaissons qu'à ce stade, cette description des réseaux des enquêtés tâtonne encore, et que l'interprétation de ces éléments n'est pas encore entreprise.

Recoder les métadonnées Facebook sur l'activité numérique

Sur les indicateurs d'activité, il nous a finalement fallu ni plus ni moins que recoder les métadonnées de Facebook... Reprenons. En Avril 2014, en ouvrant le premier export d'Algopol, nous avons constaté que l'API Facebook donnait des informations sur le volume d'activité d'un compte : le nombre de publications, de photos, de statuts, de liens. En juillet 2014, il est devenu clair d'une part qu'il était impossible de savoir à quoi correspondaient ces indicateurs (une photo postée depuis un mobile ne semble pas être comptée comme une photo postée depuis un ordinateur), et d'autre part que ces indicateurs ne permettaient pas de mener certaines analyses (par exemple distinguer les partages de liens avec ou sans message). Christophe Prieur a donc entrepris de recompter les statuts de chaque enquêté en fonction du « status_type » indiqué par Facebook : ce champ des métadonnées de Facebook précise si c'est un « *status* », « *shared_story* » (un partage), « music », etc.

Tableau 22 : *Les status_type de Facebook (volume sur les enquêtés CSA)*

985511	status
971598	link, app_created_story
128970	photo, shared_story
112212	status, mobile_status_update
108516	link
98798	status, wall_post
93937	link, shared_story
83870	status, approved_friend
51056	video, shared_story
49579	photo, added_photos
27632	photo
16821	status, app_created_story
13620	photo, tagged_in_photo
10138	swf, app_created_story
8781	question
6568	video
5648	link, approved_friend
2159	video, added_video
1680	checkin
1555	photo, mobile_status_update
345	link, created_note
301	link, created_event
287	video, app_created_story
171	offer
118	status, shared_story
47	link, created_group
39	music, shared_story
19	offer, shared_story
7	music, app_created_story
6	music
2	swf
2	photo, app_created_story

Le but était de construire cinq indicateurs d'activité pour chaque enquêté : le nombre de statuts sec, le nombre de statuts avec un lien sans message, le nombre de statuts avec un lien et un message, le nombre de statuts avec une photo sans message, le nombre de statuts avec photo et un message. Certes, les indicateurs ont été calculés … mais ces activités sont largement insuffisantes pour décrire les pratiques, et de plus les *status_type* de Facebook sont trop instables pour être fiables. Par exemple, un lien partagé peut être

compté comme « shared_story » ou non, sans que nous ne comprenions la différence d'activité pour l'enquêté. A partir de novembre 2014, Baptiste Fontaine a donc entrepris de recoder les « status_type » de Facebook en combinant les métadonnées des champs disponibles.

Pour préciser. Pour chaque enquêté, l'export Algopol construit un fichier avec une ligne par activité faite sur Facebook, par clic. Ces lignes précisent un certain nombre de champ qui sont les métadonnées Facebook de l'activité : « from_id » indique le hash de l'utilisateur qui a fait l'action, « created » le *timestamp* de la date du clic, « message » le contenu textuel du message, etc. Mais en explorant les données, il s'est avéré que le champ « message » est renseigné si l'activité est un statut, que le champ application liste aussi bien Candy Crush que « Shared_story », et que donc il fallait rentrer dans une combinaison de certaines valeurs de différents champs pour identifier l'activité. C'est essentiellement le champ « story » qui a été décortiqué. Celui-ci indique le texte que Facebook raconte pour expliquer aux amis l'activité de l'enquêté, par exemple « X a partagé un statut » ou « X a commenté la photo de Y ». Les story semble à première vue sans fin tant il existe de combinaisons possibles et d'imbrications relationnelles des activités : par exemple « 4100 aime que 70d5 ait indiqué aimé 4528 » est une combinaison de *like* d'un *like* de page. Le nombre de situations identifiées devait être réduit pour identifier non pas tous les clics possibles et imaginables, mais les clics significatifs de l'activité des enquêtés.

Nous avons ainsi formulé les règles pour identifier 92 « guessed_type », allant de la publication de statut sec à la création d'évènement, en passant par le commentaire sur la photo publiée par un alter ou du lien posté dans un groupe[1]. La liste est indiquée en annexe III-2, ainsi que leur définition[2]. Cela fait beaucoup de cas possibles, mais nous savons en fait qu'il en manque : par exemple nous n'avons jamais réussi à identifier le champ renseigné par « l'état » (ou l'humeur) dans un statut, de même que nous ne savons pas quand un commentaire mentionne un ami … Et nous savons aussi que certaines activités sont imprécises, par exemple nous verrons dans l'analyse des partages de liens que Facebook nous indique « X a partagé un lien » sans qu'il soit possible de savoir si ce partage vient de la page partagée ou d'un statut publié préalablement par un ami. Les *guessed_type* ne sont donc pas parfaits … mais ils paraissent déjà plus précis que les métadonnées Facebook pour étudier des activités réelles des internautes et notamment montrer les multiples pratiques possibles.

[1] Le projet a décidé de rendre disponible les règles utilisées pour recoder Facebook ainsi que de proposer le code développé par Baptiste Fontaine en *open source*.

[2] Et le code pour traiter les exports Facebook est disponible en licence ICC.

Les moyennes, médianes, logs, déciles… et les dires

Dans l'excel des sociologues, il y a une ligne par enquêté et pour chacun les 92 colonnes de l'activité, les 4 colonnes du formulaire « ego » de l'application, les 10 colonnes sur son réseau social, les 5 colonnes du profil Facebook, les … Nous sommes montés je crois à ~300 indicateurs par enquêtés, le nombre n'étant pas très stable puisque nous affinons régulièrement certains indicateurs et parfois ré-agrégeons ces indicateurs. Et pourtant, le volume ne fait pas la qualité. Que faire de ces 300 indicateurs ? La statistique traditionnelle s'attache à décrire chacun en termes de moyenne, médiane, écart-type. Ainsi, pour les enquêtés CSA, la moyenne du nombre d'amis par compte est de 150 amis, la médiane à 78 et l'écart-type à 311. Cet écart-type montre que la moyenne et la médiane ne sont pas appropriées. Si ces descriptions statistiques sont nécessaires, elles sont très insatisfaisantes pour deux raisons. A l'échelle d'un individu, les discontinuités de la pratique rende les moyennes peut signifiantes : si un internaute publie en moyenne 5 statuts par mois, il peut aussi bien publier exactement 5 statuts par mois, ou en publier 60 au cours d'un mois et plus aucun le reste de l'année. Il faut donc envisager d'observer la dérivée des indicateurs, et pour cela de revenir à des fichiers temporels. Ensuite à l'échelle de l'échantillon, comme dans beaucoup d'activités en ligne il semblerait que les activités sur Facebook suivent des modèles de longue traîne. Sur le nombre d'amis par exemple, très peu d'enquêtés CSA ont beaucoup d'amis (13 en ont plus de 1.000) et beaucoup d'enquêté ont « peu » d'amis (16 % de l'échantillon a 20 amis ou moins). De ce fait dire que la moyenne du nombre d'amis est à 150 n'est pas une description optimale. Une solution utilisée pour intégrer ce type de répartition dans des modèles économétriques et de prendre le « log » de l'indicateur, mais cette mesure en base 10 fait perdre un sens pratique ; une solution statistique consiste à utiliser non pas la valeur absolue des indicateurs mais la catégorisation, c'est celle que je retiendrai principalement par la suite.

Cette question sur la restitution des indicateurs est l'occasion de rappeler quelques caractéristiques de l'activité numérique Tout d'abord, un dispositif numérique est évolutif dans le temps et certains artefacts apparaissent ou disparaissent. Par exemple, si un statut publié par un enquêté le 1er avril 2008 n'a pas de *like*, ce n'est pas parce que l'enquêté n'a pas d'amis, ni parce que le statut n'est pas drôle ; c'est juste que le bouton *like* n'existait pas encore. Ce cas peut paraître anecdotique, mais il illustre les aléas qui font qu'observer une activité dans la durée est complexe. Ensuite, rien ne dit que la pratique des internautes est régulière et continue : c'est le problème de faire une moyenne de statuts par mois, alors qu'il peut y avoir des mois avec statuts et des mois sans. Les « allers-retours » des adolescents et les changements de vie des adultes (la mise en couple, le travail, le premier enfant) conduisent les internautes à faire évoluer leur pratique dans le temps, à l'échelle d'un mois comme à l'échelle de quelques années. La discontinuité des pratiques numériques brouille les pistes d'analyse … il faut soit calculer un indicateur global de continuité de l'activité d'un compte (est-ce que le nombre d'activité par mois est régulier ou varie par exemple) ; soit calculer indicateur par indicateur la période à observer pour avoir une activité régulière. Enfin, l'environnement numérique oblige aussi à revoir le sens des activités sur Facebook. En 2008, un jeune de 20 ans peut construire son réseau social sur Myspace, Skyblog, Facebook ou Twitter. En 2012, certaines plates-formes ont disparu des radars (Myspace et Skyblog), et d'autres ont vu le jour (Snapchat, Tumblr, pour ne citer qu'elles). Un adolescent en 2008 ne met

pas de musique sur Facebook puisqu'il les met sur Myspace, alors qu'un adolescent en 2012 met des liens YouTube sur Facebook puisque c'est une manière de dire ses goûts musicaux qui ne se fait pas sur Tumblr ni sur Snapchat. L'évolution de l'environnement impacte ainsi directement la description des usages et l'évolution de la plate-forme.

L'hétérogénéité des pratiques et l'instabilité du dispositif, des usages, de l'environnement obligent à ne pas donner trop vite du sens aux données. Le projet Algopol permet de décrire empiriquement les usages sur Facebook, mais il faut retourner voir les enquêtés pour comprendre le sens de ces usages. Par exemple avoir 100 amis peut être bien trop pour certains enquêtés et la honte pour d'autres. Lucie Charbonnet a pu initier le pan qualitatif de ce projet en faisant « raconter » aux enquêtés leur carte Algopol[1].

Rendre leur data aux enquêtés, livrer un dataset à la recherche ?

Une dernière remarque sur le traitement des données découle de l'idée de « sortir » des *datasets*. Deux autres vies des données sont ainsi envisagées : restituer leur data aux enquêtés personnellement et livrer un dataset anonymisé à la communauté de la recherche globalement. Pour le premier point, il s'agit d'un parti pris presque idéologique : ce sont les enquêtés qui sont propriétaires de leurs données, et non pas les plates-formes. Ce positionnement est porté par certains militants du web, à travers des projets comme Mesinfos[2] ou des mises en scène artistiques. Ainsi Albertine Meunier entreprend depuis plusieurs années de récupérer son historique de recherches sur Google et d'en faire « quelque chose », dernièrement c'était un livre[3]. Pour Algopol, l'équipe projet souhaite donc développer une fonctionnalité de l'application pour permettre aux enquêtés de télécharger le fichier de leur mur. Ce serait pousser l'expérience de *self-data* un cran plus loin et ouvrir des réutilisations personnelles des données que nous n'imaginons nous-mêmes sûrement pas[4]. Il faudrait « juste » un peu de temps pour développer cette fonction …

Pour le deuxième point, la nécessité d'anonymiser le corpus rend la concrétisation de l'idée (un peu) plus compliquée à implémenter. Livrer un *dataset* à la recherche se fait facilement sur les blogs (Leskovec, 2009) ou sur Twitter puisque les *tweets* sont publics. Sur Facebook, les sociabilités ordinaires relèvent de l'expression privée et il faut donc anonymiser les données. Une première expérience a été vécue douloureusement par le projet « Taste, Ties and Time » (T3), le *dataset* de comptes Facebook livrés à la communauté a dû être très rapidement retiré du web car plusieurs comptes avaient pu

[1] Pour l'explicitation de la démarche d'entretiens sur la carte voir http://algopol.huma-num.fr/appresults/entretien-avec-claire/, consulté le 04/06/2014.

[2] http://mesinfos.fing.org/, consulté le 03/06/2015.

[3] http://www.albertinemeunier.net/google_search_history/google_search_history.htm, consulté le 28/05/2015.

[4] A titre d'illustration des usages variés d'Algopol, une personne s'est fait tatouer tatoué son réseau …

être désanonymisés (Zimmer, 2010). Rendre les données Facebook discrètes et non identifiables représente un véritable challenge technique, pour plusieurs raisons.

Dans l'application Algopol, les données sont stockées sur le serveur sans anonymisation car les enquêtés ont besoin des noms-prénoms de leurs amis pour comprendre la carte. Pour tout export, les identifiants Facebook ou Prénom-Nom identifiés comme tel dans les métadonnées sont remplacées automatiquement par des codes, « hashés » en langage informatique[1]. Mais Facebook gère de manière hétéroclite la reconnaissance des identités : le « prénom nom » d'un individu apparait dans les données parfois sous la forme de son identifiant Facebook, parfois avec son libellé reconnu comme une entité nommé, pour ces deux cas le hash fonctionne ; mais parfois Facebook garde juste le prénom nom d'une personne, sans reconnaître que c'est un libellé... Par exemple, dans les activités qui consistent à indiquer des membres de la famille ne fonctionnent pas avec les identifiants. L'export fait alors apparaître en clair dans la *story* « 4100fd35 a indiqué que I. Bastard est sa sœur », sans que l'identification « Irène Bastard » ne soit hashée … Il n'est donc pas possible de sortir le fichier des activités des enquêtés sans risquer de laisser des identifiants dans ce *dataset*.

Une solution consisterait à retirer du *dataset* les champs *story* et message, privant les chercheurs en analyse textuelle des ressources d'Algopol. Mais même un export de ce type pourrait être désanonymisé car d'autres signatures que le nom-prénom figure dans les activités en ligne ; par exemple les dates et heures des publications permettent de retrouver un individu de manière unique dans un référentiel donné (Narayanan, A., & Shmatikov, V., 2009). Une autre approche serait de ne livrer que des informations agrégées sur chaque enquêté, par exemple un volume d'amis, un volume de statuts, un volume de … A nouveau, il est possible que la combinaison de ces volumes soit unique et permette de retrouver un individu sur Facebook. Pour ces deux dernières approches, comme la reconnaissance se fait dans un grand volume de données par des robots, il serait possible de modifier, ponctuellement, certaines informations de l'enquêté, pour que la reconnaissance échoue. Bref, anonymiser les data Facebook est un projet en soi, à date il est encore trop tôt pour identifier les indicateurs qui peuvent être livrés à la communauté sans mettre en péril contrat moral et éthique avec les enquêtés.

Conclusion sur le terrain Algopol

Que dire pour conclure ce retour sur l'expérience Algopol, à part que j'espère que les deux prochains chapitres, ainsi que les multiples autres aboutissements à venir, témoigneront que toute cette entreprise en valait bien la peine … Car entre les premiers

[1] Fonction informatique qui permet de remplacer un code par un autre sans qu'il soit possible de décoder ce nouvel identifiant.

dessins du *storyboard* de l'application et les derniers tris croisés sur le profil des enquêtés qui partagent des liens du monde.fr, il y a eu des heures de travail, de doutes, de renoncements et de renouvellements. Mais aussi des apprentissages pluridisciplinaires, des interactions loufoques sur les légumes, des témoignages saisissants sur les usages des uns et des autres.

J'insisterai sur deux points qu'il parait nécessaire de garder en tête pour lire la suite : les pratiques numériques relèvent d'une stratégie des internautes et les données ne disent pas le sens que les enquêtés donnent à ces pratiques, il convient donc de se préserver de tout déterminisme vis-à-vis des individus ; les activités numériques réelles des internautes forment un espace qui structure les objets, les documents, les plates-formes, les médias et c'est à partir de ces données qu'une description des structures de l'espace public peut, elle, être déterminée.