Statistique et société
décembre 2017

Volume 5, Numéro 3

Ouvrir les données

Sommaire
Statistique et société

Volume 5, Numéro 3

Sommaire
Statistique et Société

Volume 5, Numéro 3

Statistique et société

Magazine quadrimestriel publié par la Société française de statistique.
Le but de Statistique et société est de montrer d'une manière attrayante et qui invite à la réflexion l'utilisation pratique de la statistique dans tous les domaines de la vie, et de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

Design graphique
fastboil.net

ISSN 2269-0271

Éditorial

Emmanuel DIDIER

Rédacteur en chef de Statistique et Société

Cher Lecteur,

Pour ce numéro, nous sommes très heureux d'avoir confié l'organisation du dossier à deux jeunes docteurs, Antoine Courmont et Samuel Goëta, qui ont soutenu leur thèse l'année dernière, tous les deux sur le sujet de l'open data. Leur dossier montre que lorsqu'on dit « ouvrir les données », on ne parle pas de données préexistantes qui ne demanderaient qu'à « sortir » des institutions où elles ont été produites mais, bien au contraire, d'une activité visant à produire ces données avec les caractéristiques nécessaires à leur circulation. Définitivement, le mot de « données » est bien étrange ; on devrait dire « construits » : rien ne se donne, tout est construit ! Samuel Goëta rappelle l'histoire d'Etalab, la structure publique chargée de motiver les institutions publiques à faire l'effort nécessaire à l'ouverture ; Antoine Courmont décrit les différentes étapes nécessaires au reformatage des données ; Christian Quest explique les tensions et difficultés propres à l'établissement d'une base d'adresses ouverte ; Jonathan Gray instille une dimension internationale à la réflexion et enfin Flavie Ferchaud se demande quel type de pouvoir les données captées par les citoyens leur donnent.

Le dossier est suivi d'un article de méthode par Pierre Latouche, Charles Bouveyron, Damien Marie et Guilhem Fouetillou qui montre ce que l'analyse des tweets échangés juste avant et juste après le premier tour de l'élection présidentielle permet de dire sur le champ politique et sur l'issue du vote. Pour cela ils analysent à la fois le contenu des tweets et les métadonnées, c'est-à-dire les réseaux dans lesquels ils sont échangés.

Viennent ensuite deux articles varia. Marie Bergström nous montre tout l'avantage qu'on peut retirer à associer les enquêtes institutionnelles de l'INED sur la formation des couples aux données web captées sur le site Meetic. Enfin, Marine Billmann et Valentine Delorme font une présentation générale du mouvement du « Quantified Self ».

Bonne lecture !

DOSSIER

Ouvrir les données

Introduction

Antoine COURMONT
Chercheur post-doctorant au Centre d'études européennes et de politique comparée de Sciences Po

Initiées en France à partir de 2010, les démarches d'ouverture des données publiques ont suscité de grandes promesses de développement économique et de renouvellement démocratique. Leurs promoteurs mettaient en avant l'opportunité que constituait l'*open data* pour générer de la valeur par la création de nouveaux services, tout en accroissant la transparence administrative et les modalités de participation citoyenne. La ville de Rennes a été pionnière dans la mise en œuvre d'une politique d'*open data*, elle a été rapidement suivie par l'État et de nombreuses collectivités territoriales. Aujourd'hui, plus d'une centaine de collectivités sont engagées dans des démarches d'ouverture de leurs données, mais également de nombreuses entreprises, publiques ou privées, qui ont mis en œuvre des portails de diffusion de leurs données.

Ce mouvement devrait continuer à s'étendre dans les années à venir puisqu'une série de lois[1], votées en 2015 et 2016, imposent l'*open data* à différentes organisations. La loi pour une République numérique, dite loi Lemaire, instaure en particulier une obligation d'ouverture « par défaut », dans des standards ouverts et aisément réutilisable, des données publiques détenues par les administrations et les collectivités territoriales de plus de 3500 habitants. Parallèlement, cette loi comprend des dispositions visant à créer un « service public de la donnée » dont une des missions consiste à mettre à disposition des données de référence.

Les débats autour de ces lois se focalisent essentiellement sur les finalités de l'*open data*, passant sous silence les opérations concrètes de mise en circulation des données. Ils se basent sur le postulat que les données existent et sont des entités autonomes et détachées qui peuvent aisément être mises à disposition. Or, les démarches d'*open data* nécessitent un travail conséquent pour les institutions : les données doivent sélectionnées et préparées avant d'être ouvertes. Des infrastructures techniques doivent être développées pour faire circuler les données et assurer leur actualisation. Des ressources humaines doivent être dédiées à ces opérations indispensables mais qui restent peu visibles et souvent peu valorisées. Surtout, tout un ensemble de décisions, ayant des effets politiques sur l'usage potentiel des données, sont prises au cours des processus d'ouverture des données sans être mises en débat.

Alors que la loi Lemaire commence à être mise en œuvre, nous avons souhaité dans ce numéro interroger les opérations de mise en circulation des données à partir de plusieurs points de vue de chercheurs analysant ces démarches et d'une personnalité phare du monde des données ouvertes. Samuel Goëta revient tout d'abord sur l'importation du mouvement de l'open data au sein de l'Etat français et son institutionnalisation progressive au travers de la mission Etalab. Le dossier examine ensuite le travail nécessaire, mais souvent invisible, à la diffusion des

1. La loi Macron sur les données de transport, la loi Valter sur les données publiques, la loi de transition énergétique sur les données énergétiques et la loi pour une République numérique qui impose notamment un principe d'open data par défaut.

données. Interrogeant l'idée de données ouvertes « par défaut », Antoine Courmont pointe les différentes épreuves subies par les données avant leur ouverture. Un entretien avec Christian Quest illustre ensuite les enjeux et les difficultés de mise en œuvre d'une des bases de données de référence : la Base d'adresses nationale (BAN). Enfin, deux articles interrogent les effets des données ouvertes. Jonathan Gray présente comment les « mondes de données » (*data worlds*) sont redistribués par l'open data. Puis, Flavie Ferchaud se demande si les données produites par des organisations citoyennes ne sont pas davantage facteur d'*empowerment* que l'open data.

Mettre des données en open data, ce n'est pas ouvrir des portes derrière lesquelles seraient cachés des trésors dans lesquels il n'y aurait qu'à puiser ; c'est employer des ressources pour créer de nouveaux outils de connaissance et de transformation de la société, et pour mobiliser en même temps les publics qui s'empareront de ces outils.

ARTICLE DU DOSSIER

Une petite histoire d'Etalab : Comment l'open data s'est institutionnalisé en France

Samuel GOËTA

Co-fondateur de l'entreprise coopérative Datactivist

« Etalab » est le nom de la structure administrative qui a été créée en France en 2011 pour promouvoir l'ouverture des données publiques. Initialement tournée vers le développement de services privés innovants utilisant ces données, cette structure s'est donné ensuite d'autres objectifs : créer une communauté d'utilisateurs, transformer l'action de l'État. Parallèlement, ses critères de réussite ont évolué : au début, il s'agissait de maximiser le nombre des fichiers rendus accessibles, progressivement l'accent a été mis sur des critères de qualité, pour aboutir à la notion de « données de référence » présente dans la loi de 2016.

Les 7 et 8 décembre 2007, c'est lors d'une rencontre dans les locaux de l'éditeur O'Reilly à Sebastopol en Californie que 30 militants des libertés numériques, entrepreneurs et chercheurs ont défini ensemble huit grands principes de l'ouverture des données qui depuis définissent classiquement ce qui constitue une donnée ouverte[1]. Dix ans plus tard, la France est aujourd'hui le premier pays à avoir inscrit dans la loi un principe d'open data par défaut avec la loi pour une République numérique de 2016 qui instaure une obligation d'ouverture des données pour les institutions, les collectivités locales de plus de 3500 habitants et 50 agents ainsi que tous les acteurs investis d'une mission de service public.

Mais comment ces grands principes de l'open data ont-ils été importés en France ? Quels en ont été les passeurs ? Pour répondre à ces questions, je vais revenir sur la trajectoire de la mission Etalab qui, depuis 2011, au sein de l'Etat a été en charge de coordonner l'ouverture des données publiques en France[2]. Cet article retrace à très grands traits la trajectoire d'Etalab, une institution qui doit être replacée dans une histoire plus longue, celle de la longue émergence des droits d'accès et de réutilisation de l'accès et la réutilisation des informations publiques[3]. Retenons que c'est essentiellement par l'Union européenne que le droit à la réutilisation, c'est-à-dire à l'exploitation de données obtenues par le droit d'accès à l'information publique (loi CADA de 1978 en France) notamment pour la création de services, s'est imposé en France. Dans les années 2000, la Commission européenne a multiplié les études sur le potentiel économique de la réutilisation de l'information publique, évaluant jusqu'à 200 milliards d'euros par an la valeur de leur circulation optimale dans les pays de l'Union[4] et a adopté plusieurs directives à

1. Yu H. & Robinson D.G. (2012), The New Ambiguity of "Open Government" , *UCLA Law Review* 178, pp. 178-208.
2. Cet article résume le second chapitre de la thèse de doctorat en sociologie de l'auteur.
   Goëta S. (2016), *Instaurer des données, Instaurer des publics : une enquête sociologique dans les coulisses de l'open data*, Télécom ParisTech, Paris.
3. Ronai M. (1997), Données publiques : accès, diffusion, commercialisation, *Problèmes politiques et sociaux* 773-774, p. 68 ; Boustany J. (2013), Accès et réutilisation des données publiques. Etat des lieux en France, *Les cahiers du numérique* 9(1), 21-37 ; Trojette M. A. (2013), *Ouverture des données publiques. Les exceptions au principe de gratuité sont-elles toutes légitimes ?* Rapport au Premier Ministre.
4. Vickery G. (2011), *Review of recent studies on PSI re-use and related market*, Commission européenne.

ce sujet dont la directive Public Sector Information (PSI) de 2003 qui a fixé des règles minimales pour faciliter le droit à la réutilisation des données.

Plus de sept ans après sa création, Etalab reste une structure administrative récente aux attaches politiques et administratives instables. Nous verrons au fur et à mesure des trois grandes phases qui découpent cet article que l'action, les priorités et les équipes d'Etalab ne cessent d'évoluer témoignant de l'ancrage mouvant de l'ouverture des données dans les pratiques de l'administration.

De l'APIE à Etalab, un demi-tour en faveur de la gratuité des données

En octobre 2008, le gouvernement présentait un plan de développement de l'économie numérique pour le quinquennat, intitulé « France Numérique 2012 » qui proposait, dans son action 39, « la création d'un portail unique d'accès aux données publiques » dont l'Agence pour le Patrimoine Immatériel de l'État (APIE) devait piloter la conception. Cette agence avait été créée à la suite du rapport Lévy-Jouyet de décembre 2006 pour générer des revenus issus du « patrimoine immatériel de l'État » dont les données publiques faisaient partie. La création de ce portail s'inscrivait aussi dans le cadre de la transposition de la directive PSI de 2003. Dans le rapport dirigé par le député Franck Riester en 2010, le nom « État lab » a été proposé pour dénommer le portail que devait concevoir l'APIE pour « développer des services innovants à partir de données publiques ». Il n'est pas question alors de transparence ou de reddition des comptes aux citoyens.

Malgré ce rapport qui plaçait l'APIE au cœur du dispositif, l'agence est progressivement dessaisie du dossier au profit d'Etalab, une nouvelle structure dédiée à la diffusion gratuite des données publiques. Selon plusieurs personnes interrogées, un voyage d'étude à Washington en 2010 auquel participaient les conseillers numériques de l'exécutif ainsi que la secrétaire d'État en charge des questions numériques aurait convaincu les membres du gouvernement de demander le renvoi de l'APIE et de redéfinir le projet en faveur de la gratuité des données. D'autre part, plusieurs acteurs interrogés ont interprété l'accélération du développement de data.gouv.fr comme une volonté pour la majorité de rattraper son « retard » face aux collectivités locales d'opposition ayant engagé une politique d'open data (Rennes, Nantes, Montpellier, Paris, région PACA…). Enfin, de manière plus officieuse, Etalab aurait été créée dans l'optique de la campagne présidentielle pour faire apparaitre Nicolas Sarkozy comme un candidat « transparent ». La création de la mission Etalab s'est en effet précipitée quelques mois avant le début de la campagne officielle.

Le 30 juin 2010, le conseil de modernisation des politiques publiques, un organe interministériel en charge de la révision générale des politiques publiques (RGPP), a décidé de la création d'un « État lab », un « portail Internet recensant les données existantes et permettant leur réutilisation ». Le conseil des ministres du 24 novembre 2010 a annoncé la mise en ligne de ce portail avant la fin de l'année 2011 et a officiellement dessaisi l'APIE du dossier. « État lab » ne désigne plus le portail, devenu data.gouv.fr, mais la mission en charge de sa création. Sa direction a été attribuée à Séverin Naudet, ancien vice-président du site de partage de vidéos Dailymotion, qui a été nommé en 2007 « conseiller spécial sur Internet et le multimédia » de François Fillon. Le 21 février 2011, le décret 2011-194 a créé la mission « Etalab » placée sous l'autorité du Premier ministre et rattachée au secrétaire général du Gouvernement. Le 17 octobre 2011, Etalab a publié la Licence Ouverte qui acte de la gratuité des données publiées, autorise les usages commerciaux et impose aux réutilisateurs de citer la source.

En juin 2011, un prototype de portail a été lancé, mais data.gouv.fr n'était alors qu'une coquille vide sans les données des administrations. Le 26 mai 2011, le Premier ministre, François Fillon,

a publié un décret[5] et une circulaire[6] adressés aux ministres, secrétaires d'État et préfets. Le décret imposait que la liste des données publiques soumises à redevance soit arrêtée par décret afin de faire de la gratuité la norme à partir du 1er juillet 2012. Désignant un interlocuteur unique pour Etalab dans chaque ministère, la circulaire de mai 2011 exigeait, dans un délai d'un mois, une rencontre avec ses correspondants et dépasser le nombre de jeux de données publiés sur data.gov aux Etats-Unis et data.gov.uk au Royaume-Uni.

Figure 1 : Page d'accueil de data.gouv.fr lors de son lancement en décembre 2011

La première version de data.gouv.fr (figure 1) a été lancée le 5 décembre 2011, conformément au calendrier du projet. Etalab a particulièrement communiqué sur les 352 000 jeux de données publiés excédant les chiffres annoncés pour data.gov aux États-Unis et data.gov.uk au Royaume-Uni. Cette obsession du chiffre a conduit à découper les jeux de données de l'INSEE en un fichier par commune pour une même base afin de « gonfler » le nombre de jeux de donnés publiés. À la suite de l'élection présidentielle et de l'alternance à l'Elysée, plusieurs personnes interrogées ont douté de l'avenir de la mission Etalab et du maintien en ligne de data.gouv.fr. L'incertitude sur la politique d'open data a été exacerbée par un article publié dans Les Échos en octobre 2012 qui annonçait que certaines administrations pourraient commercialiser leurs données publiques[7]. Les membres du gouvernement avaient pourtant signé une charte de déontologie qui disposait que le gouvernement « mène une action déterminée pour la mise à disposition gratuite et commode sur Internet d'un grand nombre de données publiques[8]. » Cette incertitude rappelle que l'ouverture des données a reposé en grande partie sur les attaches politiques de la mission Etalab. En ouvrant des données, les agents ont souvent répondu à une injonction politique extraordinaire avant de participer à une pratique normale de l'administration.

5. Premier ministre. Décret n° 2011-57 du 26 mai 2011 relatif à la réutilisation des informations publiques détenues par l'État et ses établissements publics administratifs. Journal Officiel de la République Française. 27 mai 2011.
6. Premier ministre, Circulaire du 26 mai 2011 relative à la création du portail unique des informations publiques de l'État « data.gouv.fr » par la mission « Etalab » et l'application du droit de réutilisation des informations publiques. Journal Officiel de la République Française. 27 mai 2011.
7. Les Echos, « Open Data : l'État pourrait renoncer à la gratuité de certaines données publiques », http://www.lesechos.fr/journal20121017/lec2_high_tech_et_medias/0202329690871-open-data-l-etat-pourrait-renoncer-a-la-gratuite-des-donnees-publiques-501147.php, consulté le 15 décembre 2014.
8. Numérama, « Internet et l'Open Data dans la déontologie du gouvernement Ayrault », http://numerama.com/magazine/22534-internet-et-l-open-data-dans-la-deontologie-du-gouvernement-ayrault.html, consulté le 15 décembre 2014.

Vers une administration générale des données

En décembre 2012, les services du Premier ministre ont annoncé la nomination d'Henri Verdier à la tête d'Etalab. Ancien directeur de Cap Digital, pôle de compétitivité des entreprises du secteur numérique en Ile-de-France, il a aussi fondé MFG Labs, une entreprise spécialisée dans l'exploitation de données massives. Au cours de l'année 2013, de nouvelles orientations ont été données à la mission Etalab qui a changé de rattachement pour rejoindre le SGMAP (Secrétariat Général à la Modernisation de l'Action Publique). Ce changement de rattachement a eu des conséquences sur le discours des agents d'Etalab. Henri Verdier a présenté régulièrement l'ouverture des données publiques comme une manière de transformer l'action de l'État avant d'insister sur son potentiel économique comme le faisait son prédécesseur. La feuille de route du gouvernement en matière d'open data a été publiée en février 2013[9] pour donner de nouvelles orientations à la mission Etalab. Le premier changement portait sur les objectifs assignés à la mission. Plutôt que d'être évaluée sur la quantité de données publiées selon un objectif d'exhaustivité proche des préconisations des principes de Sebastopol, Etalab doit désormais publier des jeux de données « stratégiques » qui seront identifiés lors de six débats thématiques avec la société civile. A cette fin, l'équipe d'Etalab s'est assurée que la France allait améliorer sa place dans le classement international de l'ouverture des données essentielles de l'Open Knowledge Foundation, l'Open Data Index. Après avoir obtenu une 16e place en 2013, plusieurs données « essentielles » (la base de données des lois, décrets et ordonnances et les codes postaux) ce qui a hissé la France à la 3e place du classement. Le gouvernement a alors engagé une campagne de communication autour de ce résultat (figure 2).

Figure 2 : Illustration diffusée sur le compte Twitter officiel du gouvernement français suite à la publication des résultats de l'Open Data Index.

En outre, la feuille de route annonçait le lancement d'une nouvelle version du portail data.gouv.fr. Pour sa refonte, Henri Verdier déclarait s'inspirer des principes proposés dans le livre qu'il a coécrit avec Nicolas Colin, *L'Âge de la multitude*[10], dans lequel ils développent l'idée selon laquelle la richesse dans l'économie numérique dépend de capacité d'un acteur à capter la

9. Etalab, « La feuille de route du Gouvernement en matière d'ouverture et de partage des données publiques », http://www.etalab.gouv.fr/article-la-feuille-de-route-du-gouvernement-en-matiere-d-ouverture-et-de-partage-des-donnees-publiques-115767801.html, consulté le 12 décembre 2014.
10. Colin N. & Verdier H. (2012), *L'Age de la Multitude. Entreprendre et gouverner après la révolution numérique*, Paris, Armand Colin.

valeur de la « multitude. » Le portail a été remis à plat à l'issue d'une consultation lors de laquelle « l'écosystème » était invité à répondre à un questionnaire en ligne et à participer à des ateliers de « codesign. » Présenté comme un espace de mise en relation entre l'administration et les réutilisateurs, le nouveau portail permet de republier un fichier après l'avoir traité afin que l'administration puisse bénéficier du travail de la « multitude ». Lancé le 18 décembre 2013 à Matignon par le Premier ministre, le nouveau data.gouv.fr était présenté comme une plateforme qui, selon les éléments de langage, « permet de "faire vivre" les données et de rencontrer des innovateurs permettant de faire naître de nouveaux services. » Dans sa communication, Etalab insistait tout autant sur la valeur des données publiées sur data.gouv.fr que sur celle de la « communauté » qui s'y active. Cela s'est traduit dans les objectifs opérationnels assignés à la mission Etalab qui doit accroitre le nombre d'utilisateurs actifs de la plateforme et de réutilisations recensées sur data.gouv.fr sur son site, comme on peut le voir dans le projet de loi de finances 2015 (figure 3). Avec la refonte du portail, Etalab s'est donc vue attribuer de nouveaux objectifs qui ne consistent pas uniquement à ouvrir les données, mais à les « faire vivre » en s'assurant de leur réutilisation et de l'existence d'une communauté autour des données.

INDICATEUR 7.2 : Ouverture et diffusion des données publiques							
(du point de vue du citoyen)	Unité	2012 Réalisation	2013 Réalisation	2014 Prévision PAP 2014	2014 Prévision actualisée	2015 Prévision	2017 Cible
Nombre de ressources en open data (site « data.gouv.fr »)	Nombre	SO	SO	SO	36 000	37 000	40 000
Nombre de contributeurs actifs (site « data.gouv.fr »)	Nb de comptes actifs	SO	SO	SO	3 500	4 000	10 000
Nombre de réutilisations (site « data.gouv.fr »)	Nb	SO	SO	SO	1 400	2 000	5 000

Figure 3 : Indicateurs de performances de l'ouverture des données publiques dans le projet de loi de finances 2015.

Du point de vue des données, la mutation la plus importante d'Etalab concerne la création d'une fonction d'Administrateur Général des Données (AGD), traduction française du rôle de Chief Data Officer attribué dans plusieurs administrations états-uniennes. Henri Verdier a été nommé à cette fonction par décret le 19 septembre 2014. L'AGD a pour mission de coordonner l'action des administrations en matière d'« inventaire, de gouvernance, de production, de circulation et d'exploitation des données. » Dans le cadre de ses missions, il peut être saisi par tout citoyen ou toute personne morale et doit remettre chaque année un rapport au Premier ministre sur la gouvernance des données. L'équipe de l'AGD[11] a conduit plusieurs expérimentations avec des administrations notamment sur des domaines variés tels que la consommation électrique de l'État, les vols de voiture ou l'emploi. Henri Verdier a ensuite été nommé en septembre 2015 directeur interministériel du numérique et du système d'information et de communication de

11. Administrateur Général des Données : « L'équipe », https://agd.data.gouv.fr/lequipe/, consulté en juin 2016.

l'Etat (DINSIC). Laure Lucchesi, spécialiste de la transformation numérique et des stratégies fondées sur la donnée, lui a succédé à la direction d'Etalab.

De l'ouverture des données à l'ouverture des gouvernements

A partir de 2014, une autre évolution importante pour Etalab a été l'entrée de la France dans le Partenariat pour le Gouvernement Ouvert (Open Government Partnership). En avril 2014, lors de la conférence de Paris sur le gouvernement ouvert, Henri Verdier a annoncé que la France allait rejoindre cette organisation internationale créée en 2009 par le président Obama, pour promouvoir les « bonnes pratiques » en matière de transparence des États, la participation des citoyens et la collaboration avec la société civile. En plus de leurs missions liées à l'ouverture des données, les agents d'Etalab assurent une grande partie de la représentation de la France au sein de cette institution internationale et élaborent, en concertation avec la société civile, les engagements que prend le gouvernement dans le cadre de son plan d'action. Ces nouvelles missions liées à des enjeux de participation sont critiquées en particulier par l'association Regards Citoyens qui considèrent qu'elles détournent Etalab de sa mission : « il est temps qu'Etalab se recentre sur les actions concrètes simples et rapides au cœur de ses missions pour vraiment faire avancer la transparence et l'Open Data. »[12] Par la suite, la France a intensifié son engagement dans le Partenariat en rejoignant son comité directeur en août 2014 puis en prenant sa présidence en 2016. En décembre 2016, Paris a accueilli le sommet annuel du Partenariat pour un Gouvernement Ouvert, un évènement coordonné par Etalab et accueillant près de 4 000 participants de 140 nationalités venus débattre de transparence de l'action publique, la participation citoyenne et l'innovation démocratique. Conçu comme une « COP21 de la démocratie », l'évènement a connu médiatiquement un certain « désintérêt » selon un journaliste de Mediapart »[13] et un collectif d'associations de défense des libertés numériques, dont la Quadrature du net, la Ligue des droits de l'homme, l'April ou Framasoft, a annoncé qu'il ne participerait pas à ce sommet dénonçant un « leurre » du fait de la surveillance des citoyens en état d'urgence et de l'opacité de certaines décisions gouvernementales[14].

Promulguée le 7 octobre 2016, après près de quatre ans d'annonces, de consultations et de débat, la loi pour une République Numérique impose un principe d'ouverture des données par défaut à toutes les administrations et collectivités locales de plus de 3500 habitants et 50 agents. La loi crée une nouvelle mission pour Etalab dans son article 14, celle de coordonner le service public de la donnée dont la mission est de garantir la qualité et la disponibilité de « jeux de données de référence » qui présentent le plus fort impact économique et social comme la base SIRENE des entreprises (ouverte sur data.gouv.fr depuis janvier 2017), la base adresses ou encore le cadastre (ouvert depuis octobre 2017). Cette nouvelle mission rompt en quelques sortes avec l'obsession de la quantité de données des premiers temps d'Etalab pour garantir la qualité de certaines données très réutilisées.

La trajectoire d'Etalab au cours de ces sept dernières années que nous venons de parcourir souligne l'instabilité et la fragilité de la politique publique d'ouverture des données. En effet, comme on a pu le voir à travers l'alternance, l'évolution de son contexte légal et l'attribution régulière de nouveaux objectifs en matière d'open government ou de réutilisation des données, cette structure reste très liée à ses attaches politiques et à un contexte législatif mouvant. Par exemple, depuis novembre 2017, Etalab a encore changé de rattachement quittant le Secrétariat Général de la Modernisation de l'Action Publique (SGMAP) pour rejoindre la Direction

12. Regards Citoyens, « La France presidera-t-elle l'Open Communication Partnership? », https://www.regardscitoyens.org/la-france-presidera-t-elle-lopen-communication-partnership/, consulté en juin 2015.
13. Mediapart, « Henri Verdier : « Vers un développement démocratique durable » », https://www.mediapart.fr/journal/culture-idees/061216/henri-verdier-vers-un-developpement-democratique-durable?onglet=full, consulté en décembre 2017.
14. NextInpact, « À l'approche du sommet mondial de l'OGP, les consultations en ligne dans la tourmente », https://www.nextinpact.com/news/102389-a-l-approche-sommet-mondial-l-ogp-consultations-en-ligne-dans-tourmente.htm, consulté en décembre 2017.

interministérielle du numérique et du système d'information et de communication de l'Etat (DINSIC).

Conclusion

On peut retenir de la trajectoire d'Etalab que cette structure a contribué à porter l'attention sur les données publiques, bien au-delà de la question de leur ouverture. Le rapport annuel de 2015 de l'AGD l'explique bien en invitant à s'intéresser aux « données de gestion » produites dans les systèmes d'information de l'État : « La plupart des données existantes sont aujourd'hui produites dans de grands systèmes de gestion informatisés, et ne sont pas connues ni repérées comme telles. Une histoire connue dans les communautés open data concerne cette grande municipalité qui souhaitait ouvrir son portail d'open data et recherchait dans ce but des données concernant les pratiques culturelles. Il lui fallut près d'un an pour réaliser que l'application de gestion des bibliothèques municipales recelait un trésor [...] De telles données, issues des grands systèmes de gestion, représentent aujourd'hui un sujet central de la gouvernance de la donnée » Cette description des données de gestion correspond, à peu de choses près, aux « sources administratives » de la statistique publique évoquées par Desrosières[15] lorsqu'il distingue deux sources de la statistique publique : d'une part, les « enquêtes » produites spécifiquement par des institutions dédiées selon des normes scientifiques et d'autre part, les « sources administratives » issues de services « dont les activités de gestion impliquent la tenue, selon des règles générales, de fichiers ou de registres individuels, dont l'agrégation n'est qu'un sous-produit, alors que les informations individuelles en sont l'élément important, notamment pour les individus ou les entreprises concernés. » Si les sources administratives sont utilisées depuis longtemps par les statisticiens, les politiques d'open data, dont la mission Etalab a été la cheville ouvrière depuis près de sept ans, ont contribué à mettre en lumière les données brutes de l'Etat comme le matériau informationnel de la transformation et de l'ouverture du gouvernement.

15. Desrosières A. (2005), Décrire l'Etat ou explorer la société : les deux sources de la statistique publique, *Genèse* 58, 4-27.

ARTICLE DU DOSSIER

De la « donnée » à la « donnée ouverte » : les épreuves de l'ouverture des données

Antoine COURMONT

Chercheur post-doctorant, Centre d'études européennes et de politique comparée, Sciences Po[1]

L'expression « ouverture des données publiques » peut suggérer, à tort, qu'une simple décision administrative suffirait à rendre accessibles des « gisements d'information » préexistants. En réalité, c'est d'un processus d'ouverture qu'il faut parler. On peut le décomposer en trois phases : l'identification, la publicisation, et l'extraction proprement dite. Ce n'est qu'au terme des « épreuves de diffusibilité » que les données peuvent être dites « ouvertes ».

La loi pour une République numérique, dite loi Lemaire, votée en 2016, instaure un principe d'open data « par défaut » pour les administrations publiques[2]. Cela renverse le paradigme en matière de diffusion des informations publiques : l'ouverture devient la règle, la fermeture l'exception. Cette loi marque ainsi le passage d'une logique de demande à une logique d'offre. Sauf exception, les données sont considérées comme ouvertes par défaut. L'analyse ethnographique du processus de mise à disposition des données publiques au sein d'une collectivité territoriale française révèle toutefois l'impensé de ce principe d'open data par défaut : les données ne peuvent être ouvertes par défaut, puisqu'elles acquièrent précisément ces caractéristiques de diffusibilité au cours du processus d'ouverture.

Le travail de catégorisation est en effet l'enjeu central du processus d'ouverture de données. Ouvrir une donnée nécessite en premier lieu d'identifier ce qu'est une donnée, puis de déterminer sa diffusibilité. Pour comprendre comment une donnée devient une donnée ouverte, il faut observer, en situation, ce que les acteurs entendent par le terme de « données », de « données publiques », de « données ouvertes », de « données candidates », de « données personnelles », de « données sensibles », etc. Ce travail de catégorisation est au cœur de la politique d'ouverture de données : plus que d'éventuelles résistances d'acteurs accrochés à leur pouvoir, il détermine pourquoi certaines données sont, ou non, mises à disposition d'un nouveau public.

Les « barrières » qui restreindraient l'ouverture des données occupent une place centrale dans les écrits, militants ou académiques, sur l'*open data*[4]. Janssen et al. pointent le fait que l'ouverture des données implique une transformation institutionnelle d'un système fermé à un système ouvert[5]. Gray et Davies appellent à passer d'une vision de la « libération » des données

1. antoine.courmont@sciencespo.fr

2. Article 6 de la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique
3. Les résultats présentés dans cet article sont issus d'une thèse de doctorat en science politique. Courmont A. (2016), Politiques des données urbaines. Ce que l'open data fait au gouvernement urbain, Sciences Po, Paris, 423 p.
4. Martin C. (2014), Barriers to the Open Government Data Agenda: Taking a Multi-Level Perspective, *Policy & Internet*, vol. 6, no 3.
5. Janssen M., Charalabidis Y. et Zuiderwijk A. (2012), Benefits, Adoption Barriers and Myths of Open Data and Open Government, *Information Systems Management* (*ISM*), vol. 29, no 4, 258-268.

à une politique de recomposition de l'infrastructure informationnelle[6]. Les barrières à lever pour permettre ce changement sont nombreuses : héritage institutionnel, emprise politique, aversion au risque, complexité de la démarche, contraintes techniques, incertitude juridique, culture « fermée » des administrations, etc. A partir de l'étude de politiques locales d'open data, Peter Conradie et Sunil Choenni pointent trois principaux facteurs limitant la mise à disposition des données : le stockage décentralisé des données, les sources externes de données, le non-usage de la donnée dans le cœur du service public[7].

En soulignant les nécessaires recompositions institutionnelles, cette littérature pointe le fait que les données ne sont pas autonomes d'un environnement social. Néanmoins, en se focalisant sur les « barrières » de l'open data, ces auteurs s'abstiennent de s'intéresser aux données elles-mêmes. Dans leur perspective, les données préexistent à l'ouverture et elles ne jouent aucun rôle dans ce processus dans la mesure où il suffit de convaincre leurs propriétaires de lever des « barrières » pour les diffuser. Or, l'ouverture des données redéfinit tout autant les données que l'environnement dans lequel elles sont insérées. Pour comprendre comment une « donnée fermée » devient une « donnée ouverte », il est nécessaire d'adopter une perspective relationnelle qui étudie symétriquement les données et l'environnement dans lequel elles sont insérées.

Au cours du processus de diffusion, la donnée subit en une série d'épreuves qui détermine son avenir en définissant sa « diffusibilité[8] ». En catégorisant différemment la donnée, l'épreuve participe au détachement de la donnée de l'infrastructure informationnelle dans laquelle elle est insérée. En effet, la donnée n'est jamais brute, mais elle est toujours étroitement associée à un ensemble de personnes, de pratiques, de technologies, d'institutions qui les produisent, les maintiennent et les utilisent. Pour permettre sa diffusion et son utilisation dans un environnement autre, il est nécessaire de délier l'ensemble de ces attachements afin de rendre la donnée autonome de ce cadre initial. Cela oblige à reprendre une à une les composantes de l'attachement, ce qui exige de prendre en compte la donnée dans toutes ses dimensions : juridiques, techniques, économiques, politiques, etc.

Le travail de détachement est toutefois indissociable d'un travail d'attachement. Dissocier, c'est créer de nouveaux liens autant que d'en défaire d'autres. Pour susciter l'intérêt à l'*open data*, la donnée doit être associée à de nouveaux enjeux (développement économique, émergence de nouveaux services, simplification du travail des agents, image innovante de la collectivité, etc.) et à de nouveaux utilisateurs (entreprises, développeurs, citoyens, agents de collectivités territoriales, etc.). Tout autant que la donnée, le public des données ouvertes est défini au travers de ces épreuves. Des usages potentiels des données sont préfigurés tout au long du processus, un travail de cadrage, bien connu des sociologues des techniques, que Madeleine Akrich a défini comme des scripts[9]. Etudier le processus de diffusion des données c'est analyser symétriquement les médiations qui détachent et les médiations qui attachent.

Ainsi, le travail préliminaire à la mise à disposition des données éprouve à la fois le détachement des données de leur infrastructure informationnelle et leur attachement à de nouveaux utilisateurs. Multiples, ces épreuves de *diffusibilité* peuvent être regroupées en trois catégories : l'identification, la publicisation et l'extraction.

---

6. Gray J. et Davies T. (2015), Fighting Phantom Firms in the UK: From Opening Up Datasets to Reshaping Data Infrastructures ? Paper presented at the Open Data Research Symposium, Ottawa.
7. Conradie P. et Choenni S. (2014), On the barriers for local government releasing open data, *Government Information Quarterly*, vol. 31, p. S10-S17.
8. La diffusibilité est un terme scientifique caractérisant l'aptitude d'une substance fluide à se diffuser (gaz, lumière). Ce terme est préféré à celui, plus courant, de diffusabilité, utilisé pour désigner quelque chose que l'on peut diffuser. En effet, il est plus adapté à l'argument de ce chapitre puisqu'il souligne le fait que la donnée est transformée au cours du processus d'ouverture pour acquérir des propriétés de diffusibilité.
9. Akrich M. (1987), Comment décrire les objets techniques ?, *Techniques et culture*, no 9, p. 49-64.

Identifier

Identifier les données candidates à l'ouverture n'est pas une tâche aisée. La donnée ne préexiste pas à son ouverture : elle n'est pas déjà-là, prête à être mise à disposition. Comme le soulignent Jérôme Denis et Samuel Goëta, les données publiques « *ne sont pas disponibles en l'état, prêtes à être libérées. Leur existence même est loin d'être une évidence*[10] ». Une institution telle que la communauté urbaine de Lyon ne connaît pas son patrimoine de données de manière exhaustive. Il n'existe pas en son sein de catalogue[11] recensant l'ensemble des données traitées par les différents services et directions[12]. Dès lors, comme l'affirme un urbaniste du système d'information : « *aujourd'hui, on ne sait pas ce que l'on possède* ». En effet, contrairement aux applications, les données n'ont pendant longtemps pas été perçues comme un actif stratégique du système d'information.

Pour ouvrir une donnée, il faut dès lors, en premier lieu, savoir ce qu'est une donnée, quelles données existent et qui les détient. Cette étape cruciale d'identification des données est effectuée par les responsables du projet *open data*. A la manière d'explorateurs, ils s'appuient sur différents outils, qu'ils perfectionnent peu à peu, pour tenter de s'orienter au sein de l'espace informationnel de l'agglomération : des cartes, des acteurs « référents » au sein des institutions, des réseaux de producteurs de données, la connaissance d'utilisateurs externes, etc. Ces outils leur offrent différentes pistes qu'ils vont suivre de manière décousue afin d'identifier les données candidates à l'ouverture. Comme lors de toute démarche exploratoire, l'incertitude règne, et les découvertes sont parfois le fruit du hasard de la sérendipité. Ce travail exploratoire ne se déroule pas uniquement au début du projet, mais il est permanent et toujours actualisé afin de continuer à enrichir le portail de mise à disposition de données.

L'exploration de l'espace des données de l'agglomération lyonnaise repose sur une démarche pragmatique et opportuniste. Elle ne repose sur aucune stratégie déterminée préalablement. En l'absence de cadre planificateur, l'identification du patrimoine informationnel est loin d'être exhaustif. Seule une minorité de donnée est identifiée à l'issue de cette épreuve. L'épreuve d'identification peut ainsi marquer la fin du processus d'ouverture pour certaines données. C'est le cas de celles dont les porteurs de projet n'ont pas connaissance, ou de celles identifiées, mais qui ne sont pas associées à la plateforme par manque de volonté, oubli ou d'autres priorités. A l'issue de l'épreuve d'identification, certaines données sont écartées de la démarche d'ouverture de la communauté urbaine. Seule une minorité de données accèderont à l'épreuve suivante de la publicisation.

Publiciser

Etape phare de l'ouverture des données, la publicisation est l'association des données à un nouvel usage et un nouveau public d'utilisateurs. De la même manière que les ingénieurs mettent en scène des utilisateurs tout au long de la phase de conception des objets techniques[13], les producteurs préfigurent des usages potentiels à partir desquels ils jugent de l'opportunité de mettre à disposition leurs données. Ils avancent un certain nombre d'hypothèses sur les éléments qui composent le monde dans lequel la donnée ouverte doit prendre place. Ils

10. Jérôme Denis et Samuel Goëta, « La fabrique des données brutes. Le travail en coulisses de l'open data. » dans Clément Mabi, Jean-Christophe Plantin et Laurence Monnoyer-Smith (eds.), *Ouvrir,partager, réutiliser. Regards critiques sur les données numériques*, Paris, Editions de la Maison des Sciences de l'Homme, 2017, p.
11. La loi CADA impose pourtant aux administrations un catalogue de leurs informations publiques. « Les administrations qui produisent ou détiennent des informations publiques tiennent à la disposition des usagers un répertoire des principaux documents dans lesquels ces informations figurent. » (Article 17 de la loi n°78-753)
12. Ce projet de catalogage exhaustif des données de l'institution est régulièrement annoncé comme indispensable par les acteurs au sein du Grand Lyon, mais également des villes américaines étudiées (New York, Philadelphie, Chicago).
13. Akrich M. (1993), Les objets techniques et leurs utilisateurs, de la conception à l'action. dans Bernard Conein, Nicolas Dodier et Laurent Thévenot (eds.), Les objets dans l'action, Raisons Pratiques., Paris, Editions de l'EHESS, p. 35-57.

élaborent des scénarios – ou scripts – mettant en scène des acteurs et l'espace dans lequel ils vont évoluer. Ces représentations varient en fonction des producteurs et des données. Selon une direction, la mise à disposition des données peut mener à une modification des rapports de pouvoir, à une remise en cause de l'action de la collectivité, ou encore à des usages malveillants pour les citoyens. D'autres producteurs de données peuvent être davantage attentifs à des aspects économiques (crainte d'espionnage industriel), juridiques (responsabilité du producteur engagée en cas de mauvais usage), ou sécuritaire (risque pour la sécurité publique et l'intérêt général). Enfin, pour certaines données spécifiques aux processus métiers internes à la collectivité, les producteurs n'imaginent aucun usage potentiel par des acteurs externes et ne voient donc pas l'intérêt de les mettre à disposition.

Afin d'intéresser les producteurs à l'*open data*, l'enjeu est alors de pondérer ces risques perçus par les avantages potentiels à l'ouverture de leurs données. La donnée est considérée comme publiable quand le producteur estime que les avantages à associer la donnée à de nouveaux utilisateurs sont plus grands que les risques inhérents. La publicisation de la donnée consiste ainsi à représenter un « public fantôme[14] » et à le traduire en une multitude d' « êtres intermédiaires[15] », dont on ne sait s'ils seront les utilisateurs réels des données mises à disposition, mais que l'on mobilise pour déterminer le caractère de diffusibilité des données.

L'épreuve de publicisation souligne le lien étroit entre la donnée et son usage. Contrairement aux discours des militants de l'*open data* qui incitent les producteurs à diffuser leurs données sans se préoccuper des usages qui en découleront, en pratique, le processus d'ouverture des données se caractérise par la construction d'utilisateurs imaginés. Ces « êtres intermédiaires » prennent des formes multiples selon les données et les enjeux auxquels elles sont associées. Cette phase de publicisation nous révèle, dans la continuité des travaux des pragmatistes américains[16], qu'il n'existe pas un public homogène. Il y a plusieurs publics « fantômes » pour reprendre l'expression de Lippmann qui sont rassemblés autour d'un problème (*issue*) spécifique matérialisé ici par une donnée. Dès lors, dans l'ouverture des données publiques, la notion du « public » est tout autant à interroger que celle de la « donnée ». Plutôt que l'ouverture des données publiques, il faut ainsi questionner l'ouverture des données aux publics.

En corollaire de la mise en visibilité des données sur une plateforme publique, l'épreuve de publicisation révèle également les dynamiques inverses de constitution d'opacité autour des données non publiées. L'ensemble des questionnements auxquelles sont soumises les données, leur association à de nouveaux usages et utilisateurs, les incertitudes des producteurs, les scènes de négociation et d'intéressement constituent un processus de mise en visibilité tout autant qu'ils rendent invisibles les données non publicisées à l'issue de cette épreuve. Solidement attachées à leurs producteurs, ces dernières n'acquièrent pas les caractéristiques de données diffusibles.

Extraire

La dernière épreuve de diffusibilité de la donnée est son extraction. Elle correspond à l'ensemble des opérations nécessaires pour extraire la donnée de son environnement initial et la mettre à disposition sur une infrastructure de diffusion. Les données ne sont en effet jamais indépendantes d'une infrastructure technique au sein de laquelle elles sont produites et utilisées. Ces systèmes d'information n'ont pas toujours été conçus pour permettre l'extraction et la mise à disposition des données qu'ils contiennent. Une transformation de l'infrastructure

14. Lippmann W. (1927), The *phantom public*, New York, Simon & Schuster.
15. Boullier D. (2010) Le client du poste téléphonique : archéologie des êtres intermédiaires, dans *Débordements. Mélanges pour Michel Callon*., Paris, Presses de l'Ecole des Mines, p. 41-61.
16. Lippmann W., *The phantom public*, op. cit. ; Dewey J. (2010) *Le public et ses problèmes*, Paris, Gallimard.

est souvent réalisée pour permettre la mise à jour automatique des données sur le portail, un changement de format de données ou la diffusion de données volumineuses et en temps réel, tout en assurant la sécurité du système d'information de l'institution. Ces opérations mettent à l'épreuve les liens qui associent la donnée à un système technique et participent au travail de détachement de la donnée de son environnement initial et de son attachement à un système externe. Cette épreuve doit permettre de stabiliser l'identité des données comme « données ouvertes ».

A l'issue de cette troisième épreuve de diffusibilité, certaines données identifiées et publicisées ne sont pas mises à disposition du fait de contraintes techniques ou organisationnelles. Ce dernier point est particulièrement important dans la mesure où la diffusion des données est souvent une tâche qui incombe aux producteurs alors qu'elle est chronophage et peu valorisée par leur encadrement. Plus généralement, cette épreuve précise encore le public associé aux données. Certains attributs sont sélectionnés ou modifiés, des formats sont identifiés, des modes d'accès et d'actualisation sont mis en place en fonction d'usages préfigurés des données[17]. Un certain type d'utilisateur est ancré dans l'infrastructure de diffusion au travers des formats, des métadonnées, des modalités d'accès, etc.

Conclusion : de la « donnée » à la « donnée ouverte »

Au cours des différentes épreuves du processus d'ouverture, les acteurs définissent successivement ce qu'est, ou n'est pas, une donnée ouverte. L'épreuve d'identification catégorise la « donnée candidate ». La publicisation fait émerger la « donnée publiable ». Enfin, l'extraction précise ce qu'est une « donnée ouverte ». La donnée est modifiée tout au long de la chaîne de diffusion pour finir par se stabiliser comme une entité diffusable. Ce processus est réversible : des données non catégorisées comme telles peuvent le devenir, et, inversement, des données ouvertes peuvent être redéfinies comme non-diffusables.

La donnée ouverte n'est ainsi pas pré-existante à son ouverture. Au début du processus de diffusibilité, la donnée n'existe pas. Dès lors, il est impossible d'affirmer que la donnée a une essence, c'est-à-dire certaines propriétés desquelles, il serait possible de déterminer, dès le début de ce processus, sa diffusibilité. Loin d'être joué à l'avance, le processus d'ouverture des données est le résultat d'une série d'épreuves, au résultat toujours incertain, au cours desquelles les caractéristiques des données, des producteurs, des utilisateurs, sont jugées et redéfinies par les acteurs. Ces différentes épreuves de diffusibilité conduisent ainsi à recomposer le réseau des données afin qu'elles soient considérées comme « diffusables ».

17. Ce que Jérôme Denis et Samuel Goëta nomment la "brutification".
Denis J. et Goëta S. (2017), Rawification and the careful generation of open government data, *Social Studies of Science*, SAGE Publications, 47 (5), pp.604 - 629.

ARTICLE DU DOSSIER

« Mettre l'utilisateur au centre de la diffusion de données »

Entretien avec

Christian QUEST

Coordinateur de la Base Adresse Nationale au sein de la mission Etalab, président d'Open Street Map d'avril 2014 à juin 2017[1].

La loi pour une République numérique a impulsé la création d'un service public de la donnée, dont une des missions centrales est de mettre à disposition des bases de données dite de référence. Ces données « constituent une référence commune pour nommer ou identifier des produits, des services, des territoires ou des personnes ; sont réutilisées fréquemment par des personnes publiques ou privées autres que l'administration qui les détient ; leur réutilisation nécessite qu'elles soient mises à disposition avec un niveau élevé de qualité ». La base adresse nationale est l'une des neuf bases de données de référence. Dans cet entretien, nous revenons avec Christian Quest, coordinateur de la base adresse nationale au sein de la mission Etalab et président d'OpenStreetMap France de 2014 à 2017, sur l'origine de cette base et les conditions nécessaires pour que celle-ci devienne une donnée de référence.

AST[2] : Peux-tu nous présenter le projet de la base adresses nationale ?

CQ : Avant de parler des bases de données adresses, le premier problème concerne l'existence même des adresses. En France, l'adresse est gérée au niveau des communes : on en a maintenant un peu moins de 36 000, mais une très grande majorité sont toutes petites et ne sont pas outillées pour gérer des adresses, voire même n'ont jamais donné de nom de rue ni de numéro. La Poste indique que 40 % des points d'arrêts postaux, c'est à dire l'endroit où le facteur s'arrête pour distribuer du courrier, n'ont pas de numéro. Il s'agit de lieux-dits ou des hameaux où il n'y a pas d'adresses. Les maires n'ont pas d'obligation de dénommer les voies ni de les numéroter. Ils le font parce que c'est nécessaire pour la bonne organisation de la commune, pour les secours, etc. Mais, il n'y a rien qui impose l'adressage aujourd'hui. C'est un premier problème que la base adresse nationale ne va pas résoudre : une base adresse, ça ne peut répertorier que ce que l'on a nommé et numéroté.
Le deuxième problème n'est pas que l'on n'ait pas de base adresses nationale, c'est qu'il en existe plusieurs : le cadastre, la BD Adresses de l'IGN, le Répertoire des Immeubles localisés (RIL) de l'INSEE, les bases de données de La Poste, sans compter toutes les entreprises (GRDF, Enedis, Orange, etc.) qui ont des bases adresses. Beaucoup d'acteurs se sont créés des bases pour leurs propres besoins. Mais, ces bases adresses ne contiennent pas les mêmes informations. La Poste, ils ont 18 ou 19 millions d'adresses, l'IGN, 25 millions, moi, j'estime

1. OpenStreetMap (OSM) est un projet collaboratif visant à constituer une base de données géographiques libre.
2. Entretien réalisé le 21 avril 2017 par Antoine Courmont, Samuel Goëta et Timothée Gidoin (« AST »)

qu'il y en a environ 20 millions d'adresses sur le terrain en France. Finalement le problème des bases adresses en France, ce n'est pas qu'il n'y en a pas, c'est qu'il y en a trop et qu'il n'y en a aucune qui soit arrivée à un niveau de qualité qui fasse que l'on ne se pose même plus la question de savoir laquelle on utilise.

AST : Quand est venu le projet d'unifier ces différentes bases de données ?

CQ : Cela a commencé au sein d'OpenStreetMap. Nous étions régulièrement sollicités au sujet des adresses, or, nous n'avions aucune base de ce type. En même temps, nous commencions à en avoir besoin car de nombreuses données en *open data* contiennent des adresses sémantiques, mais aucune localisation géographique. Or, si nous souhaitons les remettre sur nos cartes pour ajouter par exemple les monuments historiques, il nous fallait pouvoir géocoder ces adresses sémantiques et obtenir une position géographique. Un jour, un de nos contributeurs a écrit un script qui nous a permis de récupérer les adresses à partir des plans cadastraux, et ainsi de générer une base de 16 millions d'adresses. Nous avons croisé cela avec les adresses existantes dans OSM et celles publiées en open data par certaines collectivités locales. Cela nous a permis de créer une base qui prenait le meilleur des trois que l'on a décidé d'appeler la Base Adresse Nationale Ouverte (BANO). Assez rapidement, à partir de la BANO, qui est en fait un outil de diffusion des données collectées dans OSM, on s'est dit qu'il faudrait conclure des partenariats nationaux, ou locaux, pour alimenter, mettre à jour, agréger et verser un maximum de données de qualité. Notre souhait était de mettre en place un pot commun où tout le monde vient mettre de l'adresse et tous ceux qui en ont besoin viennent se servir. C'est l'idée d'OSM, mais appliquée à une thématique unique qui est l'adresse. En faisant cela, nous avons donné un grand coup de pied dans la fourmilière : cela faisait des années que l'on entendait parler d'une base adresses nationale et qu'elle n'existait pas dans les faits.

AST : Comment expliques-tu qu'un projet qui était assez historique, d'une base adresse nationale qui rapproche les différentes bases des différentes administrations, n'ait jamais pu voir le jour, et que ce soit le côté ouvert qui ait permis d'unifier un peu tout ça ?

CQ : C'est parce qu'on est dans une logique ouverte justement. Ça dépend où on met la priorité : est-ce qu'on met la priorité sur l'intérêt des utilisateurs des données, ou est-ce qu'on met la priorité sur les producteurs des données ? Avec BANO, on a mis l'intérêt sur les utilisateurs. Ils ont besoin de quelque chose, on a de quoi leur répondre, on fait. Nous n'avons pas de contraintes financières ou de business model à défendre. Lorsque c'est le cas, malheureusement, l'intérêt du producteur tend à passer souvent devant l'intérêt des utilisateurs. Et c'est ça qui bloque la mécanique.

AST :A partir de la BANO, comment est-on arrivé au projet de la base adresse nationale (BAN) impulsée par Etalab ?

CQ : La BANO a fortement intéressé Etalab. Ça faisait des années qu'ils avaient un problème tout bête : plein de jeux de données en open data avec des adresses, et pas de possibilités simples de les géocoder, sauf en utilisant les services de Google ou similaires, qui posent d'énormes problèmes juridiques et de souveraineté. Si un Etat n'est pas capable de faire ce genre de chose soit même sans faire appel à une multinationale étrangère, c'est un peu gênant. Donc, ça a fortement intéressé Etalab, et ils m'ont recruté pour faire avancer le dossier adresse le plus loin possible.

AST : Pourquoi avoir conclu un partenariat entre l'IGN, La Poste et OSM en avril 2015 ?

CQ : En mettant en place un partenariat avec les principaux acteurs producteurs de bases adresses en particulier La Poste et l'IGN, l'objectif de la BAN est de dénombrer toutes les

adresses, savoir qu'une adresse existe. Il faut croiser et agréger ces bases de données. Ce n'est pas facile à cause du manque de circulation de la donnée entre les administrations et les différents services de l'État. Obtenir 90% des adresses, c'est assez facile. Faire les quelques % restants, c'est plus compliqué. Si chacun fait le boulot dans son silo, il parviendra à couvrir 90 voire 95%. Si par contre on travaille tous en commun, on va tous amener l'énergie qui va permettre d'atteindre les 100%.

AST : Et à l'heure actuelle, comment se fait le mélange entre les données de La Poste, les données d'OSM et les données IGN ?

CQ : Alors pour l'instant, la BAN qui est diffusée aujourd'hui, la BAN que moi j'appelle version 0, ce sont des données qui sortent de l'IGN. L'IGN agrège des données qui proviennent du cadastre, des données qui proviennent de La Poste, des données qui proviennent d'un certain nombre de partenaires, de quelques collectivités, de quelques services de pompiers, de gens qui collectent aussi des informations pour l'IGN. Il n'y a pas de données OpenStreetMap, pour une simple raison, c'est qu'il y a une incompatibilité de licence. Les données OpenStreetMap sont sous licence ODbL[3], qui implique un partage à l'identique, alors que l'IGN va commercialiser ses données sous une licence qui n'implique pas le partage à l'identique. Le problème n'est pas la commercialisation : on peut commercialiser des données OpenStreetMap, mais, même si on les commercialise, il faut les commercialiser avec une clause de partage à l'identique. On ne peut pas s'en séparer. Et cette clause n'existant pas, il est impossible de diffuser de la donnée provenant d'OpenStreetMap. Donc la BD adresses ne peut pas intégrer des données OpenStreetMap.

AST : Il y a donc deux bases adresses nationales : la BAN et la BANO ?

CQ : Oui, il reste deux bases. Alors, au niveau d'OpenStreetMap, on a aussi fait le choix de ne pas faire un cumul entre BAN et BANO. Si on faisait le mix des deux, on aurait, une base qui serait plus à jour et plus complète que la BAN, et si OpenStreetMap fait ça, en gros on tue la BAN parce que les utilisateurs attendent la base la plus complète et la plus à jour. On s'est un peu interdit de le faire pour laisser une chance à la BAN de prendre, en visant le long terme quitte à déroger à notre règle "je peux faire, je fait".

AST : Qu'est-ce qui a donc changé avec la BAN ?

CQ : Ce qui a changé avec cette convention, c'est qu'il existe une base adresses de sources officielles (IGN, La Poste), disponible sous une licence de type ODbL, qui répond aux critères de l'open data. Cela a permis à Etalab de mettre en place un géocodeur qui a bonne réputation. En 2016, 446 millions de requêtes ont été traitées, en 2017 plus d'un milliard de requêtes. Par exemple, depuis que la base SIRENE est en open data, tous les mois on a un stock mensuel qui sort, et, tous les mois, je le géocode. Beaucoup de gens sont déjà extrêmement contents et surpris qu'on fasse ça. Ça leur évite de le faire, parce que c'est quand même un boulot qui est un peu compliqué, et qui nécessite des ressources et des compétences.

AST : Quelles sont les prochaines étapes pour la BAN ?

Il faut que l'on arrive à passer à une vraie gestion collaborative des adresses. Quand je parle de collaboratif, ce n'est pas uniquement entre les signataires de la convention, c'est plus largement collaboratif avec la DGFIP et le cadastre, avec un maximum de collectivités, avec un maximum d'autres opérateurs, avec l'INSEE. Pourquoi aujourd'hui l'INSEE gère-t-elle sa propre base

3. L'Open Database License (ODbL) est une licence de style « copyleft » permettant de copier, de modifier, de faire un usage commercial sous trois conditions : citer la source, redistribuer sous des conditions de partage identiques les modifications, maintenir ouverte la base de données redistribuée.

adresses ? Je leur ai posé la question : la qualité. C'est pour des raisons de qualité qu'ils ont leurs propres données. Cela signifie que le produit disponible aujourd'hui ne convient pas. Donc, plutôt que de travailler tout seul de son côté pour maintenir les données, il faut collaborer avec les autres acteurs sur le même domaine. Ces projets coopératifs, c'est une forme de coopérative de fainéants. Si on additionne notre travail plutôt que de refaire la même chose chacun de notre côté, soit on fait globalement moins de travail, soit on obtient un meilleur résultat avec la même quantité de travail. C'est ça que j'essaye de faire comprendre. Mais ce n'est pas évident parce que la culture de la collaboration, du partage de l'information, du partage des données est très limitée voire inexistante. La culture dominante, aujourd'hui, c'est l'inverse. C'est vraiment très très compliqué de faire comprendre que si on veut arriver à fournir le meilleur service final à l'usager, et je reviens à l'usager, il faut qu'on s'allie. Ensemble, on est plus fort pour abattre les derniers problèmes.

AST : C'est cela qui permettra à la BAN de devenir véritablement une « donnée de référence »[4] ?

CQ : Pour qu'une base fasse référence, en fait il faut que qu'elle soit la plus fraîche possible. La fraîcheur des données est presque plus importante que le critère certifié d'autorité ou de qualité de la donnée. C'est vraiment la fraîcheur. Si la longueur du tuyau entre l'apparition d'une donnée sur le terrain et la sortie dans le référentiel est trop longue, qu'est-ce que font les gens ? Ils font une copie du référentiel et puis ils font les mises à jour eux-mêmes. Finalement c'est pour ça que les pompiers ont des bases adresses. Ils ne devraient pas en avoir, ils devraient récupérer la base officielle et puis terminé. Pourquoi les pompiers mettent à jour leur base adresses ? Il y a un truc qui ne va pas. Vraiment la fraîcheur de la donnée, c'est un critère primordial pour devenir une référence. Une référence ça ne s'impose pas. C'est exactement comme un standard, les gens vont converger vers quelque chose en disant, "on sait que ça marche, c'est bon, et on ne se pose pas la question". On utilise parce que l'on sait qu'on ne va pas trouver ou faire mieux.

Pour arriver à cela, il faut mettre la priorité sur l'utilisateur et non sur le producteur. Par exemple, l'INSEE est utilisateur de ses propres données, donc ils ont une logique d'utilisateur, ils produisent des données parce qu'ils les utilisent eux-mêmes. Ok, très bien, ça c'est une chose. Mais après, quand on a des utilisateurs extérieurs, il faut aussi se mettre à leur place. Dans un domaine tout autre, qui n'est pas l'adresse, mais qui est quand même indirectement lié à l'adresse, la gestion du code officiel géographique est aujourd'hui totalement inadaptée aux utilisateurs. Les changements qui interviennent dans le découpage administratif français, ce n'est pas six ou dans le meilleur des cas trois mois après leur entrée en vigueur, que l'information doit être diffusée ! C'est trois mois AVANT l'entrée en vigueur qu'elles devraient être diffusées pour que les utilisateurs puissent anticiper ces changements. À tel point que même en interne à l'INSEE, sur le code officiel géographique, il y a un code officiel géographique officieux, temporaire, qui est utilisé sans attendre le mois de mars ou avril, le code officiel officiellement publié, parce que sinon, pendant trois mois, on ne pourrait enregistrer aucune nouvelle entreprise dans la base SIRENE. En interne, il y a un Code Officieux Géographique qui est produit parce qu'ils en ont besoin sans attendre, mais les ré-utilisateurs à l'extérieur qui ont exactement les mêmes besoins attendent des mois. C'est un vrai problème. Pour certaines données de référence, il faudrait peut-être dissocier le rôle de producteur du rôle de premier utilisateur de la donnée. Sinon, il y a une logique métier dans la production des données, or, une donnée de référence ne doit pas être faite pour un métier, elle doit être faite pour tous les métiers. Il faut donc vraiment, vraiment, mettre l'usager de la donnée au centre. Et le premier usager de toutes les données de l'État, c'est l'État lui-même. Or, aujourd'hui, il y a une très mauvaise circulation des données. On s'aperçoit que le travail a été refait plusieurs fois parce qu'on ne s'échange pas les données.

4. Au sens de la loi de 2016 ; voir : https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000031366350&idArticle=LEGIARTI000033205649&dateTexte=&categorieLien=cid

ARTICLE DU DOSSIER

Quand les mondes de données sont redistribués : Open Data, infrastructures de données et démocratie

Jonathan GRAY

Chargé de cours en Études critiques des infrastructures,
Département des Humanités numériques, King's College, London

L'open data, défini comme un ensemble d'idées et de conventions qui transforment l'information en une ressource publique réutilisable, est promu pour des objectifs variés : améliorer la transparence des institutions publiques, créer des projets qui renforcent la démocratie, stimuler la croissance économique. Les infrastructures sociales et techniques qui soutiennent l'open data recomposent les « mondes de données » : de nouveaux collectifs sociaux se forment, de nouvelles pratiques créatrices de sens apparaissent. Des initiatives politiques transnationales voient le jour. Loin d'être une simple « libération » des données, cela ne va pas sans traduction, médiation, et de nouvelles pratiques sociales. Mais ce mouvement peut-il servir de base d'une délibération démocratique plus riche, ou est-il voué à institutionnaliser socialement diverses formes de bureaucratisation et de marchandisation ?

Le pouvoir apparent et le potentiel de transformation des nombres a depuis longtemps inspiré de l'émerveillement, de l'inquiétude, et des actions de diverses sortes, pour le meilleur et pour le pire. La donnée est imaginée et utilisée pour comprendre, piloter et remodeler le monde, que ce soit pour stimuler la croissance économique ou pour redistribuer les ressources ; pour exploiter les ressources naturelles ou pour conserver les écosystèmes ; pour promouvoir la santé publique et l'éducation ou pour réprimer le crime ou la dissidence ; pour transformer la terre en territoire et les personnes en citoyens, consommateurs, travailleurs, camarades ou suspects. Des comptes rendus de tels projets peuvent être trouvés dans une littérature de plus en plus fournie d'histoire et de sociologie de la quantification et des statistiques[1].

Qu'arrive-t-il à ces pratiques sociales et à ces imaginaires de la quantification quand des technologies numériques facilement disponibles facilitent la création, l'analyse et la reproduction

1. Voir par exemple :
   - Hacking I. (1985), Making People Up, in T. C. Heller, M. Sosna, & D. E. Wellbery (Eds.), *Reconstructing Individualism: Autonomy, Individuality and the Self in Western Thought* (pp. 222–236), Stanford, Stanford University Press.
   - Porter T. M. (1986), *The Rise of Statistical Thinking, 1820-1900*, Princeton, Princeton University Press.
   - Porter T. M. (1996), Trust in Numbers: The Pursuit of Objectivity in Science and Public Life, Princeton, Princeton University Press.
   - Desrosières A. (2002), The Politics of Large Numbers: A History of Statistical Reasoning, (C. Naish, Trans.), Cambridge, Harvard University Press.
   - Espeland W. N. & Stevens M. L. (2008), A Sociology of Quantification, European Journal of Sociology / Archives Européennes de Sociologie, 49(3), 401–436. https://doi.org/10.1017/S0003975609000150
   - Rottenburg R., Merry S. E., Park S.-J., & Mugler J. (Eds.) (2015), *The World of Indicators: The Making of Governmental Knowledge through*
   - *Quantification*, Cambridge, Cambridge University Press.
   - Bruno I., Jany-Catrice F., & Touchelay B. (Eds.) (2016), The Social Sciences of Quantification: From Politics of Large Numbers to Target-Driven Policies, New York: Springer.

des données par différents publics ? Quelles sortes de changements, de dynamiques, de controverses, de visions et de programmes peuvent-elles être observées quand les données deviennent numériques ? Une réponse récente à ces questions peut être trouvée dans le phénomène de l'open data qui peut être compris comme ensemble d'idées et de conventions visant à transformer l'information en une ressource publique réutilisable.

Les conventions de l'open data

Par les pratiques et initiatives de l'open data, les sous-produits de l'administration, de la gouvernance et des autres activités des institutions publiques peuvent être transformés en une ressource « brute » - décrite indifféremment comme « le nouvel or », « le nouveau pétrole », ou « le nouveau terroir ». Cet objectif est poursuivi à travers une série de conventions légales, techniques et sociales visant à rendre les données publiques de manière à catalyser diverses formes d'innovation au-delà du secteur public et à travers lui. Ces conventions tirent parti d'une constellation de cultures et de normes associées à l'open source, au logiciel libre, à la culture libre, au piratage citoyen, au journalisme de données, aux données reliées, à la méthode agile de développement informatique et aux communautés du web 2.0. Si l'on transpose l'analyse d'Howard Becker sur les conventions qui font tenir ensemble « les mondes de l'art »[2], quelles sortes de « mondes des données » ces conventions d'open data soutiennent-elles ?

Certaines de ces conventions visent à rendre les données légalement et techniquement réutilisables. Ainsi, les *licences ouvertes*, les régimes légaux et les politiques de l'information visent à atténuer les effets du droit d'auteur et des droits de propriété sur les bases de données qui peuvent entraver la réutilisation des données publiques, en stipulant clairement qu'elles peuvent être légalement réutilisées sans redevance ni autorisation. Elles tirent parti de pratiques légales associées avec le logiciel libre ou open source, la culture du libre, les groupes d'accès ouvert et de science ouverte. Il existe de fortes normes pour des formats de fichiers documentés publiquement, accessibles, structurés, qui donnent la priorité à des formats de données lisibles par les machines tels que « CSV » (« comma-separated values »), « JSON » (« JavaScript Object Notation ») et « XLS » (« feuille de classeur Excel ») par rapport à des formats de sortie imprimée comme « PDF ».

Figure 1 : Copie d'écran du portail de données « data.gov.uk » montrant un aperçu de données de dépenses provenant du cabinet du Premier Ministre du Royaume-Uni.[3]

2. Becker H. S. (1984), Art Worlds, Berkeley, University of California Press.
3. On remarque les métadonnées de la licence d'open data en dessous du titre du fichier, le bouton de téléchargement des données brutes sur la droite, et un apercu des données structures dans la moitié inférieure de la page.

D'autre part, des centaines de « portails de données » locaux, régionaux ou nationaux émanant d'états, de citoyens, d'organisations non gouvernementales et d'entreprises réunissent des données de différentes sources pour les rendre plus faciles à trouver et à réutiliser. Par exemple, la figure 1 montre une page du portail « data.gov.uk » consacrée aux données de dépenses provenant du cabinet du Premier Ministre du Royaume-Uni ; cette page montre la licence ouverte favorisant la réutilisation, ainsi qu'un aperçu des données tabulaires structurées qui peuvent être téléchargées.

De même, les « journées de programmation informatique collaborative » ou « hackathons » visent à promouvoir la réutilisation des données ouvertes ; les « bourses », les « concours », les « incubateurs » et les « labs » font la promotion de l'innovation et de la collaboration. Ces dispositifs ont permis de donner le jour à des centaines « d'applications », de « projets de données », de « fonctions interactives », de « prototypes », de « sites Internet » et de « produits et services numériques ». Ils utilisent ces données ouvertes à des fins diverses, soit en créant de nouvelles cartes, des visualisations des données, des récits à partir des données, ou des « enquêtes de données », soit en personnalisant, adaptant, racontant, filtrant et combinant l'information d'autres manières.

Un concept malléable

Ces développements ont été promus par différentes visions de ce qui pourrait advenir en ouvrant les données officielles. Certains suggèrent que cela peut améliorer la « transparence » et la « responsabilité » des institutions publiques, par exemple en créant des projets qui montrent comment les fonds publics sont dépensés. D'autres disent que les données aideront à augmenter l'efficacité du secteur public et à réduire ses coûts, d'une part en enrôlant des « auditeurs en fauteuil » qui identifieront le gaspillage et, d'autre part, en permettant et en encourageant des acteurs non-étatiques à produire des sites Internet et des services qui sans cela seraient produits sur fonds publics. D'autres soutiennent que les données publiques peuvent être utilisées pour créer des sites et des projets qui renforcent « la démocratie » - par exemple en permettant à des citoyens de contacter des institutions ou des politiciens, ou de se coordonner autour de tâches civiques ou collectives. D'autres maintiennent que les données peuvent être utilisées par de nouvelles firmes, des entreprises de technologie et des start-ups pour « créer des emplois » et « stimuler la croissance économique ». L'open data peut ainsi être compris comme un concept malléable, qui est reconfiguré pour s'aligner avec différentes conceptions des institutions publiques, des marchés et de la vie sociale[4].

Comment est-ce que l'open data formate les pratiques sociales de quantification ? Ici, nous pouvons regarder au-delà des « fichiers de données » que valorisent les défenseurs de l'open data, et nous concentrer sur les « infrastructures de données » à travers lesquelles ces fichiers sont créés. Ces dernières sont des arrangements sociotechniques qui étayent la production des données : elles consistent en des écologies relationnelles de composants logiciels, des normes de données, des méthodes, des techniques, des comités, des chercheurs, des instruments et d'autres choses[5]. Plus généralement encore, les infrastructures de données créent des « mondes de données » dont les fichiers font partie dans lesquels différents acteurs voient les choses, s'en occupent et de s'y relient de différentes manières.

L'open data permet, et promet, différentes formes de « redistribution » et de « reconfiguration »

4. Gray J. (2014), Towards a Genealogy of Open Data, Presented at the European Consortium for Political Research (ECPR) General Conference 2014, University of Glasgow. http://dx.doi.org/10.2139/ssrn.2605828
5. Bowker G. C. & Star S. L. (1998), Building Information Infrastructures for Social Worlds — The Role of Classifications and Standards, in T. Ishida (Ed.), Community Computing and Support Systems (pp. 231–248), Springer Berlin Netherlands.Heidelberg.
Bowker G. C. & Star S. L. (2000), Sorting Things Out: Classification and Its Consequences, Cambridge, MIT Press.
Bowker G. C., Baker K., Millerand F., & Ribes D. (2009), Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment, in J. Hunsinger, L. Klastrup, & M. Allen (Eds.), International Handbook of Internet Research (pp. 97–117), Springer

de ces infrastructures de données et de ces mondes de données. Cela va des visions social-démocrates de participation populaire à des institutions publiques et aux services publics jusqu'à des politiques qui cherchent à limiter le rôle de l'état en lui faisant offrir ses données « brutes » que des acteurs non étatiques puissent utiliser comme base de leurs propres produits et services[6]. En tous cas, la distribution numérique des données signifie que le nombre des personnes pouvant y accéder et les utiliser est, en principe, démultiplié afin d'inclure quiconque dispose d'une connexion Internet et des connaissances de base. Les bases de données et le stockage de données, les technologies d'analyse et de visualisation montrent que les contextes d'usage des données du secteur public peuvent s'étendre bien au-delà des statisticiens, administrateurs, gestionnaires, chercheurs et fonctionnaires qui sont impliqués dans leur production et dans leur utilisation à l'intérieur des institutions publiques. De nouveaux « styles de raisonnement » (comme le dit Hacking) et de nouvelles manières de produire du sens sont rendus possibles du fait que les données transitent depuis les rapports administratifs vers des applications mobiles et des graphiques interactifs en ligne.

Au fur et à mesure que les fichiers sont recombinés, extraits, appariés, reformatés, reconstruits, réutilisés et que différents acteurs leur donnent des sens différents, de nouveaux mondes des données émergent. La redistribution de ces mondes de données peut être comprise d'au moins trois manières[7].

De nouveaux collectifs sociaux

Premièrement, nous pourrions considérer la redistribution des mondes de données en termes de changement de composition des « mondes sociaux » de « expérience des données » et de « travail sur les données » comme une réalisation distribuée, collective. Les initiatives d'open data aspirent à provoquer des redistributions dans les collectifs associés à l'information publique, en s'adressant explicitement à de nouveaux acteurs au-delà du secteur public (qu'il s'agisse de citoyens, de groupes de la société civile, d'étudiants, de chercheurs, de journalistes, de start-ups ou d'entreprises de technologie) à travers des mécanismes comme les réseaux sociaux, les rencontres sur le web, les listes de diffusion, les « labs », les incubateurs, les hackathons, les programmes de bourses, les initiatives d'appel à la foule, les applications, les projets de données, les « sprints » de données, les « expéditions » de données, et les sites dédiés. Les portails de données invitent les utilisateurs à télécharger les données publiques et à réaliser leurs propres applications, visualisations, sites Internet et services. Les initiatives d'open data visent à exploiter la sagesse de la foule, l'expertise et l'expérience de différents acteurs en dehors de l'état, de façon à, comme l'a exprimé une initiative du gouvernement du Royaume-Uni, « Montrez nous une meilleure voie ».[8] Ainsi, nous pourrions examiner à la fois « les pratiques » et « les imaginaires » de la participation publique, de l'engagement public, de la co-création et de l'innovation autour des données publiques. De même, notre attention peut se porter sur les effets de ces mutations, des nouvelles configurations de politique économique jusqu'à des formes émergentes d'espace public, de mobilisation, de controverse sociétale et de formation de problèmes. Nous pouvons encore examiner les « imaginaires de données » et le « langage des données ». Pour finir, nous pouvons regarder quelles sortes de collectifs sociaux – les « publics de données » - sont effectivement assemblés autour des données publiques en pratique, ainsi que les politiques et le modelage des dispositifs et des procédures de l'engagement public, et le degré d'inclusion et d'exclusion de ces publics.

6. Gray J. (2014) cité
7. Gray J. (2018), Three Aspects of Data Worlds, Krisis: Journal for Contemporary Philosophy.
8. Voir : http://news.bbc.co.uk/1/hi/technology/7484131.stm

Des pratiques créatrices de sens

Deuxièmement : nous pouvons considérer la recomposition des mondes de données en termes de changement dans « la manière dont les choses sont rendues compréhensibles », au sens où elles fournissent les conditions de possibilité de l'expérience du monde permettant de le comprendre, d'interagir avec lui et de lui donner sens. Pour paraphraser Bruno Latour : changez les instruments, et vous changerez la manière dont les choses sont visibles, conservables et faisables avec les données. Dans ce cas, nous pouvons regarder les nouvelles pratiques créatrices de sens associées à l'open data, principalement l'agrégation et la combinaison de données de différentes sources. On s'intéressera alors à l'usage du « machine learning », des algorithmes et des nouvelles techniques analytiques ; ou encore aux nouvelles sortes de visualisations des données et « d'expériences des données » ; ou, pour finir, à des produits et dispositifs technologiques qui facilitent les différents modes d'entrée en relation avec les données – tels que les plateformes en ligne, les applications mobiles, la géolocalisation, le balisage, l'annotation, la production participative, les notifications en temps réel, la réalité augmentée et virtuelle, les technologies portables comme des vêtements, et les installations multimédia immersives.

Cette perspective met en évidence le fait que l'open data ne change pas seulement qui peut « utiliser » différentes ressources de représentation au sujet du monde, mais il facilite aussi de nouvelles sortes de « pratiques créatrices » pour donner sens à la vie collective. Par exemple, l'application « Walkonomics » (figure 2) combine de multiples sources de données publiques ouvertes avec des données générées par l'utilisateur pour fournir des évaluations de la « marchabilité » de différentes régions et de différents itinéraires dans plusieurs grandes villes – fournissant un nouvel outil pour collectivement quantifier, classer les lieux urbains et s'y relier.

Figure 2 : L'application « Walkonomics » pour iPhone et Android suggère la « marchabilité » de différents itinéraires en combinant différentes sources de données[9]

9. Voir : http://www.walkonomics.com/ et https://data.gov.uk/apps/walkonomics-find-walkable-route

Des projets transnationaux

Troisièmement : nous pouvons regarder comment l'open data remodèle les « projets politiques qui font le monde » (et peut-être plus récemment des projets qui « défont » le monde), comme ceux que mettent en évidence les recherches récentes sur les circuits transnationaux de la globalisation. Les projets d'open data transnationaux facilitent les nouveaux régimes de la quantification transnationale et l'agrégation, l'harmonisation et la standardisation des données.

Par exemple, « Open street map » permet aux utilisateurs d'ajouter des données géospatiales provenant d'agences publiques à des données qu'ils ont collectées eux-mêmes ou tirées de cartes1[0].De même, Le « Partenariat ouvert des marchés » crée de nouvelles normes pour que les données du secteur public soient « partageables, réutilisables, lisibles par les machines » et soient donc harmonisées et comparables à travers les frontières.[11] Autres exemples, citons le projet « Open spending » qui agrège des millions de transactions de dépenses depuis plus de 70 pays[12] ; ou encore « Open ownership » qui vise à créer un nouveau répertoire global de « qui contrôle les entreprises et qui en tire profit ».[13] Ces exemples peuvent être interprétés comme de nouveaux réseaux transnationaux d'expertise, d'échange et de transfert de connaissances permettant d'harmoniser et de standardiser le « travail des données » par-delà les frontières.

Pour conclure, j'espère que ces trois façons d'examiner la redistribution des mondes de données – en termes de collectifs sociaux, de pratiques créant du sens et de réseaux transnationaux – permettront d'attirer l'attention sur les différents aspects de la politique d'open data et d'information publique, et sur la manière dont les technologies numériques donnent lieu à différentes pratiques sociales et différents styles de quantification. Alors que la rhétorique de l'open data met souvent l'accent sur « la libération » de données préexistantes du secteur public pour alimenter l'innovation et l'extraction de valeur, nous espérons avoir montré que cela ne va pas sans traduction, médiation, et de nouvelles sortes de pratiques sociales et de mondes sociaux. Le passage des données du secteur public à divers autres acteurs donne naissance à de nouvelles sortes de « mondes de données ».

D'autres mondes possibles ?

Un mouvement qui a commencé sa vie en appelant à l'ouverture des « fichiers » officiels peut encore contenir les germes d'un programme politique plus ambitieux visant à ouvrir « l'espace public, l'imagination, la participation, la délibération, la contestation et la créativité » autour de la fabrication des données[14]. Ce qui pourrait être lu en termes d'une « institutionnalisation » populaire de certains styles de raisonnement, modes d'épreuve et genres de problèmes de quantification (hérités comme un sous-produit de l'administration et de la gouvernance du secteur public) pourrait cependant servir de base d'une délibération démocratique plus riche et significative autour de « quelles choses » sont prises en compte à travers les données, « comment », et avec quels effets. En un temps où, dit-on, la confiance du public dans les institutions décroit, et où pourtant l'échelle des grands problèmes collectifs auxquels nous faisons face est considérée comme sans précédent dans les temps modernes, de telles expérimentations publiques autour du rôle des données dans les sociétés démocratiques doivent sûrement être bien accueillies. Si l'étude de l'open data suggère sans aucun doute les façons dont ce développement peut être utilisé comme un moyen d'accélérer et d'institutionnaliser socialement différentes formes de marchandisation et de bureaucratisation, elle peut aussi, parfois, nous récompenser avec des rappels que d'autres mondes des données sont possibles.

10. Voir : https://www.openstreetmap.org/
11. Voir : https://www.open-contracting.org/about/
12. Voir : https://openspending.org/
13. Voir : http://openownership.org/
14. Gray J. (2016), Datafication and Democracy: Recalibrating Digital Information Systems to Address Societal Interests, *Juncture*, 23(3). Téléchargeable depuis http://www.ippr.org/juncture/datafication-and-democracy

ATELIER DU DOSSIER

La captation citoyenne de données urbaines favorise-t-elle l'*empowerment* ?

Flavie FERCHAUD
Doctorante, université Rennes 2

Le droit d'accès à l'information publique est devenu une norme. Les politiques d'ouverture des données publiques sont encouragées, au niveau de l'État comme des collectivités territoriales. Elles répondent à deux enjeux : la création de valeur d'une part, la transparence d'autre part. En 2014, dans un article intitulé « L'open data peut-il (encore) servir les citoyens ? », Samuel Goëta et Clément Mabi dressent un bilan nuancé, mais concluent que les données ouvertes ne créent pas d'*empowerment*. Les données en créent-elles davantage quand elles sont captées, puis partagées par des habitants ?

Au travers d'une enquête sur les lieux d'expérimentations et de fabrication numérique (fablabs, hackerspaces...), l'article analyse deux projets de captation et de partage de données sur la qualité de l'air extérieur pour enclencher une réflexion au prisme des notions d'innovation sociale et d'*empowerment*[1]. L'innovation sociale est une intervention initiée par des acteurs sociaux, pour répondre à une aspiration, subvenir à un besoin, apporter une solution ou profiter d'une opportunité d'action afin de modifier des relations sociales, de transformer un cadre d'action ou de proposer de nouvelles orientations culturelles[2]. Par définition, l'innovation est nouvelle et marque un changement par rapport à des actions et des modes de pensées[3]. Elle rejoint sur ce point celle de l'*empowerment*, qui souligne une démarche de transformation sociale. Le terme *empowerment* implique une démarche d'autoréalisation et d'émancipation des individus, de reconnaissance des groupes et de transformation sociale[4]. Il renvoie à la fois au pouvoir et au processus d'apprentissage pour y accéder. En quoi les projets « Adem » et « Ambassad'Air », qui visent à développer des outils de captation de données urbaines ayant vocation à être ouvertes, s'inscrivent-ils dans cette démarche ? La combinaison du caractère *Do-It-Yourself* (DIY - faire soi-même) des outils et ouvert des données laisse supposer un *empowerment* plus fort que dans le cas des données ouvertes par les pouvoirs publics.

Adem, un projet d'innovation sociale à la démarche de changement limitée

Timelab est une association en charge d'un des fablabs de Gand. Dans le cadre des activités du fablab, des temps de réflexion collective y sont organisés. Au cours d'un entretien, Evi, salariée

1. Goëta S. et Mabi C. (2014), L'open data peut-il (encore) servir les citoyens ?, *Mouvements*, 79, 81-91.
2. Klein J-L., Laville J-L., et Moulaert F.(2014), *L'innovation sociale*, Paris, Éres.
3. Fraisse L. (2015), Entretien avec Adalbert Evers, « Analyser en contexte la dimension normative de l'innovation sociale », *Sociologies pratiques*, 31, 15-18
4. Bacqué M-H.et Biewener C. (2015), *L'empowerment, une pratique émancipatrice ?*, Paris, Editions La Découverte/Poche.

de Timelab, raconte qu'en 2014, à l'issue d'un de ces temps, émerge le projet de fabrication de capteurs mobiles de données sur la qualité de l'air, Adem (« souffle » en flamand). La captation mobile de données sur la qualité de l'air est innovante par rapport aux capteurs fixes des organismes agrées. Le caractère novateur du projet se situe aussi sur le plan social. Comme nous avons pu le constater à l'occasion des deux réunions auxquelles nous avons participé, le projet est porté collectivement par une dizaine de personnes, des bénévoles membres du fablab accompagnés de salariés de l'association. Deux objectifs sont poursuivis : la mobilisation des cyclistes de la ville pour mesurer et publier des données sur la qualité de l'air de leurs itinéraires d'une part ; la sensibilisation à la qualité de l'air pour amener les gantois à se déplacer davantage à vélo ou à pied en suivant des itinéraires moins pollués que d'autres, d'autre part.

> « *Ici, les cyclistes sont nombreux mais il y a toujours trop de voitures… Il faut que les gens soient incités à prendre plus leur vélo, mais je connais des parents qui ont peur que leur enfant respire trop l'air pollué à vélo… On sait que l'air est pollué mais on ne sait pas où exactement. Avec Adem, tu vois, ils sauront où passer pour échapper aux tronçons trop pollués. Puis le but c'est aussi que les gens lâchent leur voiture, tu vois là on va faire une campagne de communication pour qu'ils répondent à un questionnaire sur leurs déplacements et après, ils pourront dire s'ils veulent participer à la captation des données. On va coller des stickers avec des phrases punchy.* »[5]

À ce stade du projet[6], Adem vient conforter, plus que transformer, l'action publique urbaine locale. D'abord, la pratique du vélo est favorisée depuis des années à Gand, comme en témoignent, par exemple, le prêt gratuit de vélos pour les étudiants, le réseau de pistes cyclables, les zones de stationnement des vélos, etc. Ensuite, le développement d'Adem est en phase avec la mise en œuvre du nouveau plan de déplacements urbains, présenté au printemps 2016, qui limite la place des automobiles dans le centre-ville et favorise des formes de mobilités alternatives aux déplacements en voiture. En outre, la municipalité, informée du projet Adem, n'a pas attendu la publication des données pour concevoir un nouveau plan de circulation des cyclistes. La démarche de changement est plus net sur le plan des données. Le Global Open Data Index[7] classe la Belgique première sur l'ouverture des données sur la qualité de l'air. Cependant, la responsabilité de la mesure et de la publication de ces données incombe aux régions. L'organisme flamand ne publie pas ces données en suivant les principes de l'open data[8]. En effet, les mesures sont visualisées sur Internet en temps réel sur une carte à l'échelle de la Flandre[9]. Il est impossible de zoomer pour obtenir des mesures à l'échelle de Gand. Cartographiées, les données sont découpées en dix classes : les mesures précises des polluants sont donc indisponibles. Venant pallier ces manques, Adem interroge l'action publique régionale en matière d'ouverture des données sur la qualité de l'air.

Ambassad'Air, un outil de la participation publique

Comme à Gand, la qualité de l'air est moyenne à Rennes et le sujet y fait régulièrement l'objet de l'actualité médiatique et politique, au rythme des pics de pollution. En 2015, la baisse de la limitation de la vitesse sur la rocade fait débat. C'est dans ce contexte que la Ville de Rennes sollicite l'association Bug, en charge d'un des « fablabs » de la ville, pour réaliser une étude sur un projet de captation citoyenne de données environnementales dans le cadre de la préfiguration

5. Extrait d'un échange informel (notes du carnet de terrain) avec Arno, salarié de Timelab, en juin 2016 à la suite d'une des réunions sur l'avancement du projet Adem.
6. En juin 2016, la captation des données n'avait pas encore débuté. Le développement de l'outil est seulement au stade du prototypage.
7. Le Global Open Data Index (https://index.okfn.org/) évalue le niveau d'ouverture de données jugées essentielles. La qualité de l'air est une de ces données.
8. Plusieurs définitions de l'open data ont été formulées au cours de la dernière décennie. Samuel Goëta (2016) résume dans sa thèse les trois demandes essentielles des auteurs de ces définitions : la diffusion volontaire et proactive des données produites par les agents de l'État, leur ouverture juridique et technique, leur publication sous une forme brute. Goëta S. (2016), *Instaurer des données, instaurer des publics : une enquête sociologique dans les coulisses de l'open data*, thèse de doctorat en sociologie, Télécom ParisTech.
http://en.vmm.be/air

d'un projet financé par l'Agence de l'Environnement et la Maîtrise de l'Énergie (ADEME). En 2016, le pilotage du projet est confié à la Maison de la Consommation et de l'Environnement (MCE). Avec l'aide d'autres acteurs associatifs, 16 habitants de deux quartiers prioritaires sont mobilisés de façon volontaire pour tester des capteurs de la qualité de l'air extérieur. L'enjeu est en premier lieu de faire participer les habitants à la mesure de ces données pour les sensibiliser à la qualité de l'air. La fabrication des outils de captation est placée en second plan.

> « *Alors le but premier c'est vraiment d'acculturer les Rennais à la question de la qualité de l'air extérieur. Et donc qu'ils s'intéressent à ce sujet via un outil qu'il a chez lui donc il se sent vraiment acteur du truc. […]. Donc le but, c'est une sensibilité plus forte des Rennais sur le sujet parce que le constat aujourd'hui c'est que ben… Les Rennais, la qualité de l'air ça… Ils ne connaissent pas, ils ne comprennent pas bien. […] Donc le sujet premier pour les élus c'est : non, il faut arrêter le déni, il faudrait que les habitants s'intéressent à ça, qu'ils y soient plus sensibles, qu'ils connaissent un peu mieux. C'est un outil qui permet de les impliquer parce que du coup ils sont acteurs de la connaissance et on l'espère, du coup, de l'action. C'est plus un sujet vraiment sanitaire qu'un sujet technique et finalement l'aspect technique intervient de façon secondaire et est un peu l'alibi pour mettre les habitants dans la boucle.* »[10]

Ce faisant, trois éléments du projet sont porteurs de tensions : la précision des données, source d'inquiétude pour Air Breizh, organisme agréé par l'État, qui mesure déjà la qualité de l'air à Rennes (et en publie les données en ligne) ; le caractère propriétaire des capteurs ; le rôle de la MCE et des acteurs associatifs. En effet, la MCE héberge Gulliver, une association qui rassemble des utilisateurs des logiciels libres et open source[11]. Ses membres s'opposent à l'utilisation d'outils propriétaires pour capter des données publiques et dénoncent cette dimension du projet dans des e-mails échangés sur des listes de diffusion auxquelles nous avons eu accès. Par sa nature, le projet intègre ces acteurs dans une configuration d'action collective tendue, comme en témoigne cet extrait d'entretien.

> « *Ils [services de Rennes Métropole] ont vite fait d'oublier un acteur, surtout que de certains échanges que j'ai pu voir, clairement il était dit : c'est pas du rôle de la MCE de s'impliquer ou de même donner un avis sur la conception. J'ai jamais vu ça dans ma vie. En gros, la conception c'est pas du rôle des associatifs, les associatifs ça sera pour le volet communication, diffusion auprès du grand public. C'est du rôle technique, les services numériques, SIG, on va suivre ça, il y a déjà les big data, Wi6Lab [entreprise qui développe des capteurs], ils vont faire leur truc, allez-yq! Il y a beaucoup de jeu, que j'avais pas forcément perçu là dedans, de jeu… Pas de pouvoir mais… que certains acteurs estiment que tel autre n'est pas en capacité de… ou est en capacité plus de nier ou de retarder que de faire avancer le projet donc du coup volontairement, il y a des oublis ou il y a des façons d'esquiver pour que certains acteurs ne puissent pas mettre le pied dans la porte. Clairement, ça existe. Et du coup, nous là-dedans on peut être un peu… Comment dire… Instrumentalisés, quasiment*. »[12]

La gestion de cette situation aboutit finalement à des orientations allant à l'encontre de l'*empowerment*. D'abord, la MCE achète des capteurs développés en open source, mais déjà fabriqués. Les membres de Gulliver sont écartés : ils travaillent à l'amélioration de ces capteurs en parallèle de la captation des habitants. En matière de données, le projet paraît ensuite moins porteur que d'autres dispositifs accompagnant les politiques d'open data. Alors que les concours visent ainsi à développer des services pour créer la réutilisation des données ouvertes, ce n'est

10. Idem.
11. Un logiciel libre respecte obligatoirement quatre libertés (utilisation, étude, modification, duplication), qui sont au fondement de la General Public Licence (GPL, 1989), une licence générale, applicable à tout logiciel libre et dont R. Stallman est à l'origine. En anglais, logiciel libre se dit free software, ce qui introduit une confusion avec la gratuité, peu favorable aux affaires. Pour tenter de mettre fin à cette confusion et favoriser la pénétration du logiciel libre dans le monde de l'entreprise, l'expression open source fut forgée et défendue par E. Raymond en 1998 (Broca, 2013, p. 62). Concrètement, les critères de l'open source sont moins restrictifs et injonctifs que ceux du logiciel libre.
12. Entretien réalisé en mai 2016 avec le coordinateur du projet Ambassad'Air à la MCE.

pas encouragé dans le cadre d'Ambassad'Air. Les mesures sont rendues publiques, mais « à vocation pédagogique » seulement en raison de « la marge d'erreur non négligeable »[13] qu'elles contiennent. Si cela calme les inquiétudes d'Air Breizh, la portée du projet sur l'*empowerment* s'en voit réduite.

> « *F (enquêtrice) : Et ces capteurs, ils vont produire des données qui seront aussi fiables et aussi précises que celles d'Air Breizh ?*
> *J : Non. Non, justement. On a bien ressenti, parce que pour le coup j'étais à cette réunion là, je l'ai ressenti comme ça, une inquiétude d'Air Breizh mais qui était légitime, je trouve, de dire qu'aujourd'hui, les seuls agréés pour faire de la mesure, c'est eux. Et c'est vrai, ils ont un agrément ministériel, c'est les seuls agréés pour faire de la mesure. Forcément, ils voient débouler une collectivité avec des associations qui leur disent : « non mais vous êtes gentils Air Breizh mais avec vos deux capteurs sur la ville de Rennes, vous êtes gentils mais ça va pas… Nous on va multiplier, cent capteurs, on va mesurer partout ». Il y avait de quoi déstabiliser la structure. Donc au début, ils étaient plus que réticents et avaient à dire qu'à leur connaissance, la mesure avec ce genre de capteurs était de toute façon inefficace et complètement inutile. Que ça allait donner des données aberrantes et que ça risquait d'être contre-productif puisque peut-être des gens allaient s'inquiéter outre mesure ou à l'inverse, dire : « mais regardez, il n'y a pas de pollution ! » Donc risquer de les mettre dans une situation, eux, inconfortable.* »[14]

Au final, le projet apparaît difficilement à même de créer de l'*empowerment* : les capteurs utilisés par les habitants n'ont pas été fabriqués par eux-mêmes et les données produites sont ouvertes mais à vocation pédagogique seulement. Les tensions, résultant de l'intégration d'acteurs nouveaux dans une configuration d'action collective, permettent cependant de supposer l'élargissement du spectre (DIY, caractère libre et/ou open source des technologies utilisées, rôle des militants, place de l'expérimentation…) des questions relatives à la place des données dans le gouvernement des villes.

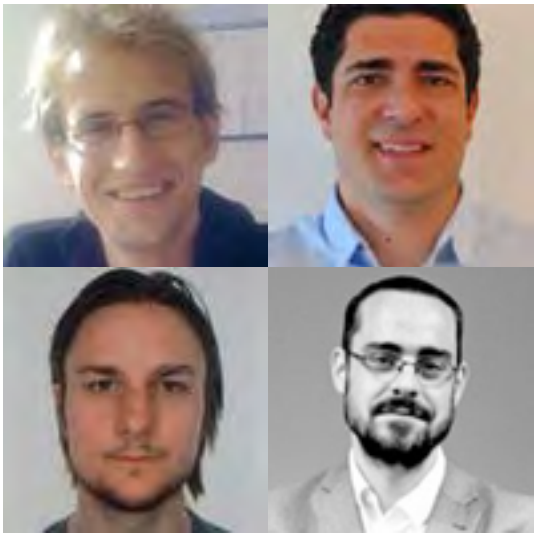
Conclusion

La réflexion porte sur deux observations seulement et les deux projets évoqués sont en cours de développement. Qu'en conclure à ce stade ? Les contextes d'émergence de ces projets sont différents mais ils sont tous deux en phase avec une action publique urbaine locale visant à sensibiliser les habitants à la qualité de l'air, à réduire la place de l'automobile, à favoriser les modes de déplacements alternatifs à l'automobile et à améliorer la qualité de l'air. On peut supposer que d'autres cas nous permettraient d'arriver à la même conclusion : la tendance à favoriser les mobilités douces et réduire la place de la voiture dans les villes européennes est forte. Concernant les politiques d'open data, les deux cas divergent. Adem vient questionner la politique régionale d'open data en matière de qualité de l'air tandis que l'ouverture de ces données ne revêt qu'un caractère pédagogique à Rennes. La transformation de l'action publique est peut-être davantage à chercher dans les processus que dans les finalités. Sur ce point, le cas d'Ambassad'Air s'inscrit dans la continuité d'une politique de participation publique de l'offre[15] à laquelle les rennais sont habitués. Le cas d'Adem permet lui d'étayer l'hypothèse selon laquelle l'entrée par la fabrication numérique peut amener les acteurs à s'emparer d'autres sujets, tels que l'open data, la qualité de l'air ou les mobilités. L'étude du projet sur une durée plus longue permettra de déceler si ses porteurs, le « fablab » au premier chef, sont à même de prendre part à la fabrique urbaine ; par exemple dans la conception de futurs itinéraires dédiés aux déplacements alternatifs à l'automobile.

12. Entretien réalisé en mai 2016 avec le coordinateur du projet Ambassad'Air à la MCE.
13. http://www.wiki-rennes.fr/Ambassad'Air#cite_note-1
14. Entretien réalisé en mai 2016 avec le coordinateur du projet Ambassad'Air à la MCE.
15. Gourgues G. (2013), Avant-propos : penser la participation publique comme une politique de l'offre, une hypothèse heuristique, *Quaderni* 79 5-12

MÉTHODES

Présidentielle 2017 : l'analyse des tweets renseigne sur les recompositions politiques



Pierre LATOUCHE
Maître de Conférences en Mathématiques Appliquées, Université Paris 1

Charles BOUVEYRON
Professeur de Mathématiques Appliquées, Université Côte d'Azur

DAMIEN MARIE
Ingénieur, Société d'accélération de transfert technologique « IDFINNOV »

GUILHEM FOUETILLOU
Professeur associé, Sciences Po Paris

Voici un exemple d'utilisation de Big Data pour observer la société. C'est d'analyse politique qu'il s'agit ici. Les auteurs tirent parti des tweets émis juste avant et juste après le premier tour de l'élection présidentielle française d'avril 2017 pour regrouper les comptes twitter, en tenant compte à la fois des contenus des messages et des liens entre ces comptes. Ayant attribué des noms de partis politiques aux groupes issus de leur travail, ils peuvent proposer une analyse des forces en présence avant et après l'élection, ainsi que des transferts entre ces forces. Les détenteurs d'un compte twitter ne forment certainement pas un échantillon représentatif de toute la population électorale : et pourtant les résultats sont remarquablement proches de ceux de l'élection.

Introduction

Emmanuel Macron a été élu à la présidence de la République sur un programme dont une des priorités est la recomposition de la vie politique. La période précédant les législatives était donc sujette à de fortes interrogations quant à la réorganisation à venir des partis politiques. Afin d'apporter un éclairage sur ce point, nous avons étudié pendant les semaines qui ont précédé le second tour de l'élection présidentielle les mouvements et transferts entre les partis, avec un prisme particulier : celui du web social. En partenariat avec l'entreprise Linkfluence, nous avons analysé la recomposition des partis sur Twitter suite au premier tour.

Figure 1 : Visualisation d'un partitionnement (« clustering ») du réseau de tweets sur le premier tour de l'élection présidentielle 2017[1].

La plupart des outils permettant d'analyser ce type de données voient les tweets comme un ensemble de documents et ont pour objectif d'étudier le choix des mots, les thèmes de discussion majoritaires, et les sentiments relayés par les tweets. Cependant, les tweets sont par nature des données plus riches que de simples documents puisqu'ils caractérisent des interactions entre des individus. Par exemple, un individu A interagit avec B s'il retweete un message de B ou s'il écrit un message faisant référence à B. Un ensemble de tweets est alors vu comme un réseau « social ». Malheureusement, les outils traditionnels d'analyse de réseaux sont eux aussi limités et ne peuvent généralement gérer que des interactions binaires (interagit ou n'interagit pas) entre les individus.

Méthode et données

L'analyse des réseaux est un domaine de recherche particulièrement actif dont un des objectifs est l'extraction automatique d'informations pertinentes à partir des interactions observées entre des individus. Les méthodes ont été créées à l'origine en sciences sociales. Depuis, l'immense majorité des outils ont été proposés par des physiciens/informaticiens afin de maximiser un critère bien particulier, la modularité. Ce critère vise à identifier des groupes d'individus ayant plus de connexions entre eux qu'avec des individus d'autres groupes. C'est le principe de la communauté. Nous observons des communautés dans les réseaux sociaux vérifiant le principe de transitivité, i.e. l'ami de mon ami est mon ami. Malheureusement, les réseaux en général et sociaux en particulier sont souvent construits à partir d'autres types de groupes. Il existe par exemple des individus ayant une forte influence sur les avis/comportements des autres. On parle alors de groupes d'influenceurs et d'influencés. De la même manière, nous trouvons également régulièrement des structures inversées où il existe plus de connexions entre des individus de groupes différents qu'entre des individus d'un même groupe. La recherche en

1. Chaque point (=nœud) représente un point d'origine d'un tweet ; le « cluster » auquel ce point appartient, identifié par une couleur, correspond à la proximité par rapport à un des candidats, identifiée sur son identifiant et/ ou son contenu. Les courbes connectant les points (=arêtes) correspondent aux échanges (réponses aux tweets). La couleur et l'épaisseur d'une arête correspondent à l'intensité (multiplicité) des échanges sur un point de discussion et sa direction (intra- ou inter- cluster). Les couleurs des arêtes ont une signification différente des couleurs des nœuds et reflètent les sujets abordés.

Mathématiques, et en particulier en Statistique, a fourni ces quinze dernières années plusieurs solutions permettant de pallier les limites des outils existants. Ces approches permettent en particulier d'identifier des individus organisés en communautés, mais également en d'autres types d'organisations sociales. La recherche française en Statistique a largement contribué aux avancées théoriques et méthodologiques dans ce domaine.

Dans le cadre d'un projet de collaboration entre les laboratoires de Mathématiques des universités Paris 1 Panthéon-Sorbonne et Paris Descartes, nous avons proposé un nouveau modèle statistique, dénommé STBM (Stochastic Topic Block Model)[2], et une méthode d'estimation associée permettant de réaliser une analyse conjointe d'un réseau et d'un ensemble de textes. Le réseau social à analyser n'est alors plus vu comme un objet binaire. Un individu A interagit avec un individu B sur un texte donné. A peut par exemple envoyer plusieurs e-mails à B. Dans ce cas, l'interaction de A vers B est caractérisée par cet ensemble d'e-mails. Pour des données de type tweet, une interaction de A vers B rassemble tous les tweets écrits par A faisant directement ou indirectement (retweet) référence à B. L'analyse de ce réseau social permet alors d'identifier des groupes d'individus en fonction de à qui ils s'adressent et de quoi ils parlent. La méthode détermine les thèmes de discussion propres aux échanges entre les groupes. Elle permet ainsi de dire : le groupe G1 identifié discute beaucoup avec le groupe G2, sur le sujet S1 identifié.

Adapté aux réseaux de taille modérée à grande (de quelques centaines à plusieurs centaines de milliers d'individus), STBM peut ainsi analyser des échanges de textes, que ce soient des e-mails, des contenus scientifiques, des tweets, etc. D'un point de vue plus technique, le modèle au cœur de l'algorithme STBM est une généralisation de deux modèles statistiques reconnus : le SBM[3] (Stochastic Block Model) qui permet de modéliser la structure d'un réseau par partitionnement (« clustering ») et le LDA[4] (Latent Dirichlet Allocation) qui permet d'analyser les thèmes abordés dans des textes. La dépendance entre les deux modèles est faite au niveau des groupes des individus : les paramètres gérant la partie du modèle liée au texte dépendent des groupes des émetteurs et récepteurs des communications. L'inférence de ce modèle statistique repose sur un algorithme CVEM (Classification Variational Expectation-Maximization) qui optimise séquentiellement la vraisemblance des parties réseaux et textes.
STBM est ainsi capable d'étudier conjointement le contenu des échanges et les interactions entre des individus ou des groupes d'individus. A titre d'exemple, STBM a été appliqué à l'analyse du réseau des e-mails de l'entreprise Enron[5], qui a connu une faillite très médiatique au début des années 2000, et à l'analyse de réseaux de co-publications scientifiques. Notons que notre plateforme Linkage.fr permet à chacun de faire traiter par STBM ses propres données de réseaux (e-mails, PubMed, Arxiv, Twitter, ...). L'exemple suivant[6] permet d'illustrer l'approche.
A partir de tous les tweets des français liés à la politique, nous nous sommes concentrés sur deux périodes : 17-18 avril et 24-25 avril 2017, c'est à dire quelques jours avant, et juste après le premier tour. Les tweets liés à l'élection présidentielle ont été extraits et formatés par Linkfluence. L'ensemble de données fourni par Linkfluence s'appuie sur Radarly, le logiciel propriétaire développé par l'entreprise permettant de suivre en temps réel la quasi totalité du web social au niveau mondial. Dans ce cas précis, la totalité des mentions des 5 candidats principaux ont été captées sur le réseau social Twitter. Ainsi, environ 5 millions de verbatims ont été extraits pour l'analyse. La méthodologie statistique a été appliquée sur les réseaux

2. C. Bouveyron, P. Latouche and R. Zreik, *The Stochastic Topic Block Model for the Clustering of Networks with Textual Edges, Statistics and Computing*, 2017 (https://doi.org/10.1007/s11222-016-9713-7).
3. K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association, 96(455):1077–1087, 2001.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3:993–1022, 2003.
5. Voir https://linkage.fr/blog/Enron-Scandal pour une analyse détaillée.
6. L'analyse et ses graphiques dynamiques sont accessibles sur le site linkage.fr, permettant de compléter l'information statique présentée dans cet article. L'accès sur le site Linkage.fr est libre, moyennant l'ouverture totalement libre et gratuite d'un compte. Rechercher : "French Twitter Politics Discussion Groups before the 2nd round of the presidential elections of 2017" sous l'onglet "Jobs".

ainsi constitués et a identifié cinq thèmes de discussion et dix groupes d'individus, sur les deux périodes (c'est-à-dire avant et après le premier tour de l'élection).

Résultats de l'analyse

Pour la 1ère période (17-18 avril), quatre des thèmes trouvés correspondent aux tweets des français à propos des principaux candidats. Cependant, il est particulièrement intéressant de constater que le cinquième thème rassemble uniquement les tweets critiquant le système politique en général. Ce thème, au cœur de la campagne, est relayé par tous les partis politiques. Un examen des comptes présents dans chacun des groupes identifiés par la méthode nous a également permis d'étiqueter chaque groupe vis-à-vis de sa tendance politique. Un groupe dont les identifiants mentionnent explicitement un parti ou un candidat donné de son parti est étiqueté du nom de ce parti. Contrairement à tous les partis, le parti socialiste se retrouve isolé et n'interagit pas ou peu avec le groupe central en gris sur la Figure 2, rassemblant les comptes Twitter des candidats et des principaux médias.

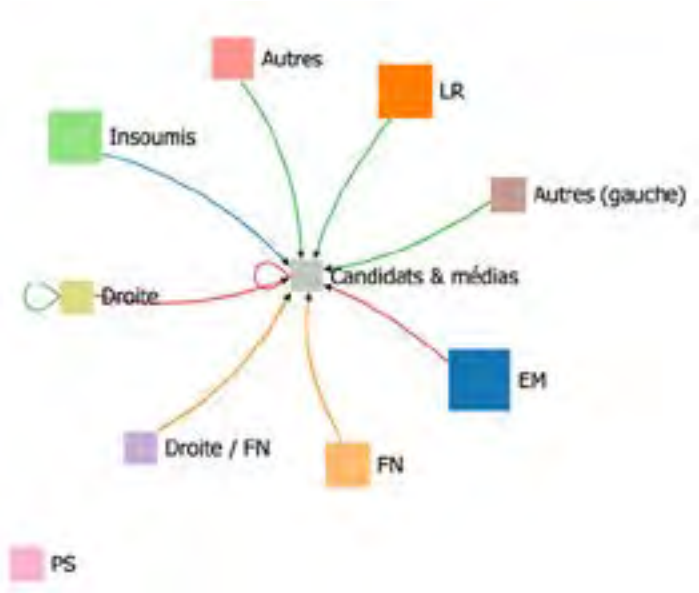


Figure 2 : représentation agrégée de la Twittosphère politique française des 17 et 18 avril[7].

De manière surprenante, les poids des partis que nous avons identifiés se sont avérés proches du vote des Français (Figure 3). 24,1% des comptes analysés ont ainsi été classés dans le groupe EM. Pour rappel, Emmanuel Macron a obtenu 24.01% des voix.

7. Chaque carré (=nœud) caractérise un ensemble de tweets regroupé sur leur proximité avec un parti ou un candidat de ce parti à partir de leur identifiant et/ ou de leur contenu. La taille de chaque nœud est proportionnelle au nombre d'éléments qu'il contient. Sa couleur a été attribuée arbitrairement pour permettre de les distinguer. Le positionnement de ces nœuds, figé sur cette vision statique, n'a pas de signification particulière ici. Les flèches indiquent les directions des tweets en terme de contenu émanant d'un groupe destiné à un ou des éléments d'un ou plusieurs groupes ou vers lui même (cf. boucles au niveau des carrés intitulés « Droite » ou « Candidats & Médias »). La couleur des flèches indique les thèmes majoritaires de discussion. Le choix automatisé de cette couleur par le logiciel est indépendant du choix des couleurs pour les nœuds, et répond à la légende suivante : thème Insoumis (bleu), thème FN (orange), thème Critique du système (vert), thème EM (rouge).

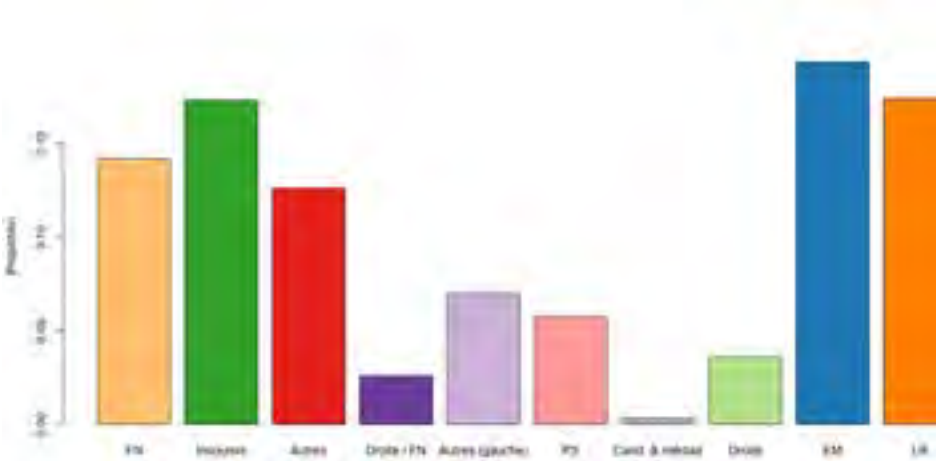


Figure 3 : poids des partis politiques sur Twitter les 17 et 18 avril[8].

Nous avons réalisé une analyse similaire sur la période 24-25 avril 2017, entre les deux tours de l'élection présidentielle, afin notamment d'observer la recomposition du paysage politique sur le réseau Twitter après les résultats du 1er tour (Figure 4). Deux thèmes sont associés à EM. Un est uniquement dédié à EM alors qu'un autre rassemble des discussions mentionnant à la fois EM et les Insoumis. Un thème correspond au FN et nous retrouvons deux thèmes de critique dont un de rejet du système politique.

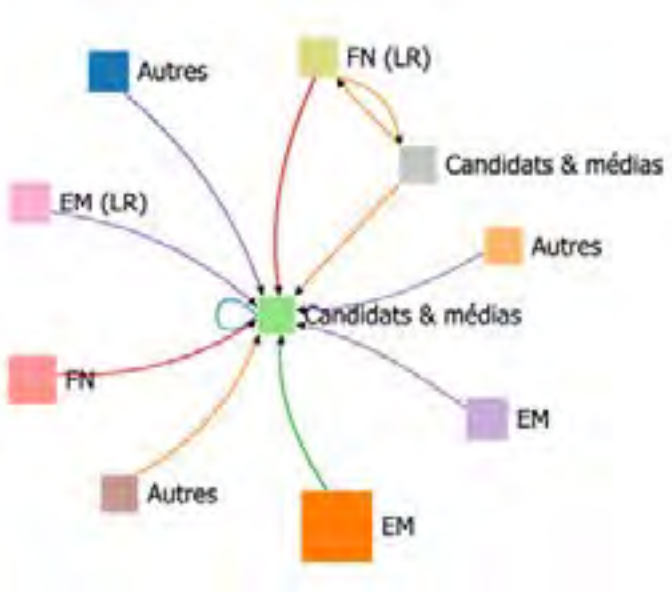


Figure 4 : représentation agrégée de la Twittosphère politique française des 24 et 25 avril.[9]

Comme pour le premier tour, nous avons pu identifier le poids des partis sur Twitter: 66% pour EM et 34% pour le FN. A la vue des résultats du 2nd tour, cette estimation du poids des partis sur le web social est bien sûr troublante. Il est néanmoins important de garder à l'esprit que le web social ne peut pas être directement utilisé aujourd'hui comme source pour le sondage car une grande partie de la population française n'est pas présente sur ces réseaux.

8. Les couleurs correspondent ici aux couleurs des carrés (nœuds) de la figure précédente.
9. Chaque nœud caractérise un groupe et sa taille est proportionnelle au nombre d'individus qu'il contient. Des couleurs leur sont attribuées sans que ce choix ni la position du nœud dans la figure n'aient d'autre but que de les individualiser. Les couleurs des flèches indiquent également les thèmes majoritaires des discussions : thème FN (rouge), thème EM-Insoumis (vert), thème EM (bleu), thème Critique du système (orange), thème Critique générale(violet). Les couleurs des flèches sont arbitraires, sans correspondance avec les couleurs des nœuds.

Fait unique, notre étude nous a permis de suivre les changements de comportement des comptes entre les deux tours. En utilisant les résultats des analyses sur les deux périodes, il nous a ainsi été possible d'estimer la recomposition du paysage politique à l'issue du 1er tour. La figure 5 permet de visualiser cette recomposition.

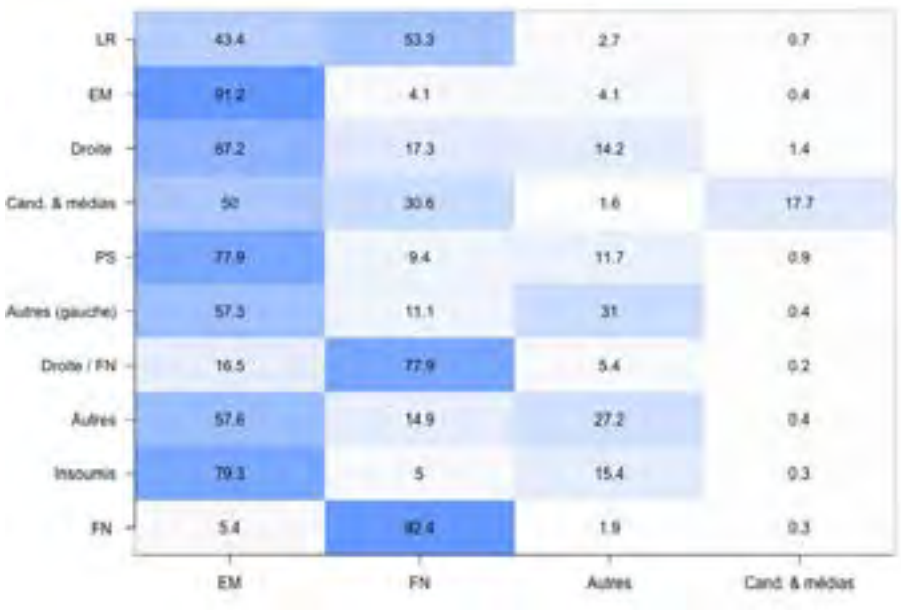


Figure 5 : Report des voix estimé par la méthode après le premier tour.
Les groupes identifiés pour les 17 et 18 avril sont en ligne.
Les groupes identifiés pour les 24 et 25 avril sont en colonne[10].

Nous avons communiqué ces résultats avant le second tour[11]. Il nous paraissait important de montrer que, sur le web social, les Insoumis semblaient finalement se tourner vers EM. Nous voulions également témoigner de la fracture que nous avions observée à droite. Une part importante des comptes actifs et proches de François Fillon a été classée FN suite au premier tour. Le reste des comptes de droite et issus de LR sont allés majoritairement vers EM.

Conclusion

Cette étude a permis de mettre en œuvre la méthodologie statistique STBM d'analyse de réseaux avec arêtes textuelles à une problématique importante qu'est l'étude d'une élection présidentielle sur le web social. Cette étude a été rendue possible par l'implémentation de la méthode STBM sur la plateforme Linkage.fr et la collaboration avec l'entreprise Linkfluence qui a capté et pré-traité les données Twitter. Outre la description synthétique de l'événement sur le web social, le résultat le plus important de cette étude est certainement la quantification de la recomposition des partis politiques entre les deux tours. En effet, la comparaison des résultats de « clustering » des deux périodes nous a permis d'estimer cette recomposition et ainsi de valider ou invalider certaines hypothèses émises par les analystes politiques.

10. Les chiffres indiquent, pour chaque colonne, les reports de voix pour le 2ème tour à l'intérieur de chaque groupe (ligne) identifié au premier tour. Le total de chaque ligne fait donc 100.
11. Tweet du vendredi 5 mai : https://twitter.com/latouche_pierre/status/860471570220929024

MÉTHODES

Quand l'Ined rencontre Meetic

Marie BERGSTRÖM

Chargée de recherche à l'Institut national d'études démographiques

Certains thuriféraires des Big Data prédisent que ces données nouvelles vont supplanter les sources traditionnelles, telles que les enquêtes sur échantillon. Des chercheurs mettent au contraire en garde contre les données massives non structurées, en attirant l'attention sur leur manque de représentativité et sur leur instabilité. Rares sont les études qui utilisent les deux types de sources. C'est le cas de ce travail, qui conclut à leur complémentarité.

Le recours aux services de rencontres sur Internet est aujourd'hui une pratique fréquente[1]. Contrairement aux annonces et aux agences matrimoniales dont l'usage est toujours resté marginal[2], les rencontres en ligne sont devenues courantes chez les jeunes comme les moins jeunes. C'est le cas en France où 12 % des femmes et 16 % des hommes, âgés de 26 à 65 ans, s'étaient déjà inscrits sur un tel site en 2013[3]. Parmi les personnes qui n'étaient pas en couple au moment de l'enquête, le taux d'usage était plus important : 25 % des femmes et 28 % des hommes[4].

La diffusion de ces services leur vaut désormais une attention des chercheurs en sciences sociales. Ces sites enregistrent de nombreuses informations qui, lorsqu'extraites, codées et analysées par le chercheur, peuvent apporter des enseignements précieux sur la formation des couples. Les données issues des sites de rencontres ont ainsi été mobilisées dans plusieurs recherches récentes consacrées notamment à l'homogamie sociale et ethno-raciale[5]. Cette nouvelle source bouscule quelque peu la sociologie du couple qui repose traditionnellement sur l'exploitation d'enquêtes par questionnaire. En France, on pense notamment aux enquêtes de l'Ined dont *Le choix du conjoint*, réalisée par Alain Girard en 1959[6] et *La formation des couples*, conduite par Michel Bozon et François Héran dans les années 1980[7]. Ces enquêtes ont prouvé leur capacité d'objectivation des rencontres amoureuses. Elles comportent cependant des limites face auxquelles les données issues des sites s'avèrent très complémentaires.

D'une part, ces sites permettent d'observer les *rencontres en train de se faire* plutôt que d'étudier les *couples déjà constitués* comme c'est le cas des enquêtes par questionnaire. Les sites

1. L'auteure remercie la société *Meetic* qui lui a fourni une partie des données mobilisées dans cet article. Elle tient aussi à remercier le programme OxPo (Oxford – Sciences Po) pour le soutien financier qu'il lui a accordé l'année où la recherche a été réalisée. Ce texte résume les résultat principaux d'un article à paraître fin 2018 dans la *Revue française de sociologie*.

2. Bozon M. et Héran F. (1987), La découverte du conjoint. I. Évolution et morphologie des scènes de rencontre, *Population*, 42-6, 943-985.
3. Bergström M. (2016), Sites de rencontres : qui les utilise en France ? Qui y trouve son conjoint ?, *Population & Sociétés*, 530, 1-4.
4. Source : enquête *Épic*, Ined-Insee, 2013-2014.
5. Skopek J., Schulz F. et Blossfeld H-P. (2011), Who Contacts Whom? Educational Homophily in Online Mate Selection, *European Sociological Review*, 27-2, 180-195 ; Lin K-H. et Lundquist J. (2013), Mate selection in cyberspace: The intersection of race, gender, and education, *American Journal of Sociology*, 119-1, 183-215 ; Potârcă G. et Mills M. (2015), Racial Preferences in Online Dating across European Countries, European Sociological Review, 31-3, 326-341 ; Schmitz A. (2016), *The Structure of Digital Partner Choice. A Bourdieusian perspective*, New-York, Springer.
6. Girard A. (2012 [1964]), *Le choix du conjoint. Une enquête psycho-sociologique en France*, Paris, Armand Colin.
7. Bozon M. et Héran F. (2006), *La formation du couple. Textes essentiels pour la sociologie de la famille*, Paris, La Découverte.

enregistrent en effet les comportements de contact des usagers et permettent ainsi d'observer le processus d'appariement des partenaires. Ce faisant, ils offrent aux sociologues du couple une opportunité méthodologique inédite : alors que les enquêtes ne captent que la *sélection* amoureuse, les sites enregistrent aussi les comportements de *refus* et d'*élimination* qui font partie intégrante de la séduction mais dont les enquêtes sont sans trace.
D'autre part, les sites de rencontres captent des *pratiques* tandis que les enquêtes enregistrent des *déclarations de pratiques*. Le questionnaire suppose une réflexivité importante de la part des enquêtés et pose la question de la mise en cohérence des discours. Dans le cas des rencontres, on peut s'interroger tout à la fois sur la capacité, la volonté et les manières – socialement et sexuellement différenciées – de raconter « son histoire amoureuse ». Les données issues des sites de rencontres font, elles, l'économie de la déclaration pour autoriser, de façon originale, une *observation quantifiée* des pratiques affectives.
À ce titre, les sites de rencontres sont un bon exemple de l'intérêt que peuvent avoir les « données massives » pour la recherche en sciences sociales et leur complémentarité par rapport aux données d'enquête. On propose d'illustrer quelques-unes de ces opportunités méthodologiques et empiriques. Pour ce faire, on présente un exemple tiré d'une recherche sur l'écart d'âge entre partenaires hétérosexuels. L'étude confronte des données d'enquête (*Épic*) à des données d'un site (*Meetic*) qui donnent des réponses différentes à la question de comment l'écart d'âge advient.

Un nouveau site d'observation des rencontres

Dans les couples hétérosexuels, l'homme est souvent plus âgé que sa conjointe. En France en 2012, les unions cohabitantes étaient caractérisées par une différence d'âge moyenne de 2,5 en faveur de l'homme[8]. Cette asymétrie des âges s'observe – à des degrés variables – dans la quasi-totalité des pays et des époques connus[9]. Elle est un objet classique de la sociologie du couple, abordée principalement à partir d'enquêtes. Celles-ci informent avec précision sur l'ampleur et l'évolution de l'écart d'âge entre conjoints mais peinent à rendre compte de la manière dont cet écart se *produit*. Est-ce que ce sont les femmes, ou les hommes, ou les deux sexes qui désirent cette asymétrie ?
Pour tenter de répondre à cette question on mobilise des données issues du site *Meetic.fr*, obtenues grâce à un partenariat avec la société éditrice du site, *Meetic France*. Deux types d'informations ont plus précisément fait l'objet d'analyses. D'une part, on a travaillé sur des données anonymisées relatives aux *profils d'utilisateurs*. Il s'agit des informations que les usagers ont renseigné au sujet d'eux-mêmes et de leurs préférences (à l'exception des pseudonymes et des photographies), comme par exemple leur sexe, leur âge et leurs préférences d'âge. Une deuxième source d'information concerne les emails envoyés sur le site. En aucun cas on n'a eu accès au contenu des messages. Seulement les métadonnées relatives aux échanges ont été analysées : les identifiants de l'expéditeur et du destinataire ainsi que la date et l'heure de l'envoi. Recoupées avec les données de profil, ces informations permettent de savoir « qui contacte qui », notamment en fonction de l'âge. Pour cette recherche on se base, d'une part, sur un échantillon constitué de l'ensemble des profils d'utilisateurs enregistrés sur le site en 2014 et ayant envoyé au moins un mail dans l'année (environ 400 000 profils) et, d'autre part, un échantillon constitué de l'ensemble des emails échangés sur le site en 2014 (plus de 25 millions d'emails).
Si ces données « interactives » ont l'avantage de donner une image dynamique des rencontres, elles comportent de nombreuses limites. Une réserve importante concerne la véracité des informations. Sur Internet, on ne donne pas toujours son âge. On arrondit. C'est le cas sur *Meetic* où les utilisateurs renseignent leur date de naissance avec une inclination pour les nombres « cinquièmes » tels que 1975, 1980, 1985 etc. Cet ajustement de l'âge, qui consiste le plus souvent

8. Daguet F. (2016), De plus en plus de couples dans lesquels l'homme est plus jeune que la femme, *Insee première*, 1613, 1-4.
9. Mignot J-F. (2010) L'écart d'âge entre conjoints, *Revue française de sociologie*, 51-2, 281-320.

à se rajeunir, se pratique par les deux sexes au même titre : le degré de surreprésentation aux chiffres « ronds » est sensiblement le même pour les femmes et les hommes[10]. Il est aussi relativement limité. Contrairement à une idée reçue, et comme le montre plusieurs recherches[11], les écarts à la vérité sont certes *courants* sur Internet mais ils sont peu *importants*. Sur *Meetic* par exemple, les utilisateurs sont curieusement plus grands et plus sveltes que la population dans son ensemble, mais les écarts à la moyenne nationale sont assez faibles : environ 2 cm en plus pour les deux sexes, de même que 2 kg en moins pour les hommes et 5 kg en moins pour les femmes. Les chiffres « ronds » ont donc un effet centrifuge sur les nombres proches sans trop distordre la moyenne, ce qui veut dire que les ajustements se situent souvent dans les bornes de +/- 4. Il y a de bonnes raisons de penser qu'il en va de même pour l'âge. Cette tendance à arrondir l'âge n'implique pas moins d'interpréter les résultats avec prudence. Plutôt que de s'intéresser à un âge donné, il s'agit d'établir les tendances d'ensemble quant à la direction et à la variation des préférences d'âge telles que manifestées sur Internet.

Ces données « massives » issues du site *Meetic* sont pour partie comparées à la dernière enquête en date sur la formation des couples en France. Cette étude, intitulée *Étude des parcours individuels et conjugaux* (*Épic*) a été conduite par l'Ined et l'Insee en 2013-2014, sous la coordination de Wilfried Rault et Arnaud Régnier-Loilier[12]. Au total, 7 809 personnes âgées de 26 à 65 ans ont répondu à cette enquête qui comportait notamment des questions sur les préférences d'âge des répondants. Comme nous le verrons, ces préférences déclarées lors d'une enquête ne sont pas les mêmes que celles déclarées *in situ* sur le site *Meetic*.

Préférence d'âge : ce que l'on dit et ce que l'on fait

Pour comprendre la manière dont l'écart d'âge se produit, l'enquête *Épic* a cherché à connaître les attitudes des femmes et des hommes envers cette caractéristique courante des couples. Il a été demandé aux répondants s'ils auraient « accepté facilement l'idée d'être avec quelqu'un qui aurait été plus jeune que vous, de 5 ans ou plus » ou « plus âgé que vous, de 5 ans ou plus ». Reprise de l'enquête sur la *Formation des couples* (1983-1984), la question permet de savoir si l'un des deux sexes tient plus que l'autre à ce que l'homme soit plus âgé dans les couples.

Le résultat est sans équivoque : en 2013 comme 30 ans auparavant, ce sont en premier lieu les femmes qui se disent attachées à un écart d'âge en faveur du partenaire masculin. Tandis que les hommes sont très nombreux à dire qu'ils accepteraient facilement une femme *plus âgée* – c'est le cas de quatre hommes sur cinq (79 %) – seulement un peu plus de la moitié des femmes s'imaginent avec un partenaire *plus jeune* (53 %). Cette réticence féminine envers un partenaire cadet est particulièrement forte chez les jeunes : parmi les 26-30 ans, seulement un tiers des femmes accepteraient facilement cette idée (33 %). Chez les hommes, les attitudes varient peu : à tous les âges, ils sont très majoritairement ouverts au scénario aussi bien d'une conjointe plus jeune que plus âgée.

Sur *Meetic*, les choses sont assez différentes. Invités à renseigner leurs préférences d'âge sur le site (c'est une information obligatoire au vue de l'inscription), les utilisateurs indiquent une fourchette entre 18 et 99 ans. En comparant l'âge renseigné par les usagers avec leurs préférences quant aux âges *minimum* et *maximum*, il est possible de mesurer autrement l'attitude des femmes et des hommes envers l'âge du partenaire. Cette analyse tranche avec les résultats d'enquête. C'est ce que montre la figure 1 en comparant la part de répondants à l'enquête *Épic* ayant déclaré facilement accepter l'idée d'un écart d'âge inhabituel en faveur de la femme (de 5 ans ou plus) avec la part d'utilisateurs de *Meetic* ayant indiqué des préférences ouvertes à un tel écart.

10. Bergström M. (2015), L'âge et ses usages sexués sur les sites de rencontres en France (années 2000), *Clio. Femmes, Genre, Histoire*, 2-42, 125-146.
11. Toma C., Hancock J. et Ellison N. (2008), Separating Fact From Fiction: An Examination of Deceptive Self-Presentation in Online Dating Profiles, *Personality and Social Psychology Bulletin*, 34-8, 1023-1036 ; Schmitz A., Sachse-Thürer S., Zillmann D. et Blossfeld H-P. (2011), Myths and facts about online mate choice. Contemporary beliefs and empirical findings, *Zeitschrift für Familienforschung*, 23-3, 358-381 ; Zillmann D., Schmitz A. et Blossfeld H-P. (2011), Lügner haben kurze Beine: zum Zusammenhang unwahrer Selbstdarstellung und partnerschaftlicher Chancen im Online-Dating, Zeitschrift für Familienforschung, 23-3, 291-318.
12. Rault W. et Régnier-Loilier A. (2015), La première vie en couple : évolutions récentes, *Population & Sociétés*, 521, 1-4.

Le décalage est apparent. Les utilisateurs de *Meetic* se montrent moins ouverts à un écart d'âge inhabituel que ne le font les répondants à l'enquête – c'est vrai pour les deux sexes. Or, ce décalage est bien plus important pour les hommes. Tandis que dans la situation d'enquête, les hommes se disaient – contrairement aux femmes – relativement indifférents à l'âge de leur partenaire, cette indifférence disparaît sur Internet. Une majorité d'utilisateurs sont certes ouverts aux femmes aînées jusqu'à 40 ans, mais après cet âge, c'est beaucoup moins le cas. Les déclarations des femmes sont, elles, plus cohérentes. Le taux d'acceptation d'un écart d'âge inhabituel est relativement similaire dans les deux situations : le coefficient de corrélation est de 0,73 contre seulement 0,24 pour les hommes.
Pourquoi ce décalage chez les hommes que l'on n'observe pas au même titre chez les femmes ? Une explication possible réside dans l'inégale réflexivité des deux sexes quant à leurs préférences amoureuses et sexuelles. Alors que les femmes sont socialisées et habituées depuis leur plus jeune âge à parler des relations affectives[13], c'est moins le cas des hommes qui manifestent plus souvent une réticence à évoquer leur vie intime, notamment dans le cadre d'une enquête[14]. Peut-être est-ce d'abord cette socialisation différenciée aux discours sur l'intimité que captent les enquêtes. Cela voudrait dire que les réponses des femmes, plus tranchées, ne reflètent pas tant une attitude plus *intransigeante* vis-à-vis de l'âge du partenaire mais avant tout des préférences plus *conscientes*. À l'inverse, l'indifférence manifestée par les hommes concernerait moins l'âge des partenaires que la *question* en tant que telle, à laquelle ils peinent peut-être à répondre. En tous les cas, *Meetic* indique que les hommes, plus que les femmes, « ne font pas toujours ce qu'ils disent qu'ils font » et interroge à ce titre les enquêtes sur la conjugalité et plus précisément la comparabilité des réponses données par les deux sexes.

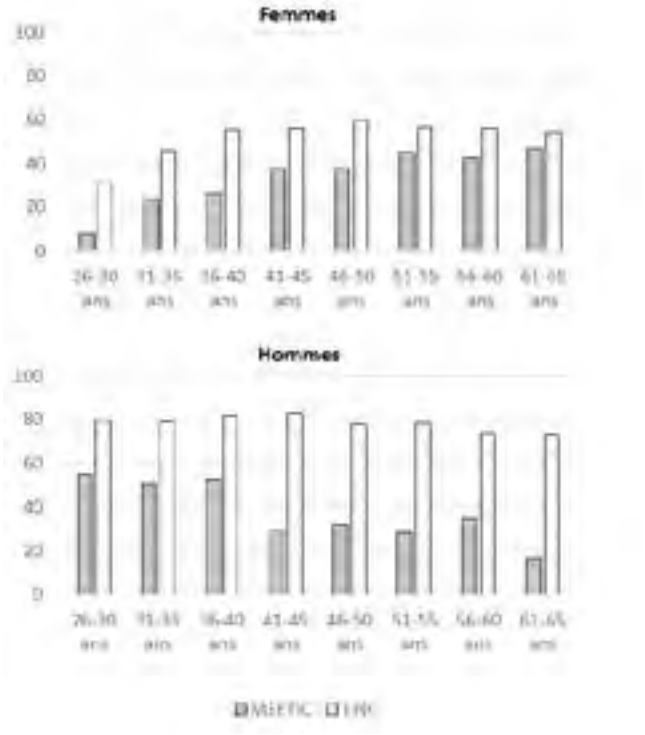


Figure 1. Part de personnes qui acceptent un écart d'âge (de 5 ans ou plus) en faveur de la femme – comparaison de réponses à l'enquête *Épic* et sur le site *Meetic* (%)[15]

13. Monnot C. (2009), *Petites filles d'aujourd'hui, l'apprentissage de la féminité*, Paris, Autrement.
14. Blin O. (1997), Paroles d'amour, mots du cœur ou vues de l'esprit ?, Mana. Revue de sociologie et d'anthropologie, 3, 57-72 ; Duret P. (1999), *Les jeunes et l'identité masculine*, Paris, Presses universitaires de France
15. **Lecture :** lors de l'enquête *Épic*, 83 % des hommes âgés de 41 à 45 ont affirmé qu'ils auraient « facilement accepté l'idée d'être avec quelqu'un qui aurait été plus âgé que [eux], de 5 ans ou plus ». Sur *Meetic*, 29 % des profils présentant des hommes de même âge ont affiché des préférences ouvertes à une partenaire plus âgée, de 5 ans ou plus.
Champ : compte d'utilisateurs actifs enregistrés sur Meetic en 2014 ; personnes vivant en France métropolitaine.
Source : base d'utilisateurs de *Meetic.fr*, Meetic Group, 2014 ; enquête *Épic*, Ined-Insee, 2013-2014.

Une histoire d'amour est une histoire de compromis

L'analyse des comportements de contact sur *Meetic* permet d'aller plus loin. Elle fait apparaître un inversement des attitudes envers l'écart d'âge avec l'âge propre des utilisateurs. Les jeunes femmes se montrent – sur *Meetic* comme ailleurs – très réticentes envers un partenaire plus jeune, mais avec l'âge les utilisatrices s'ouvrent manifestement à cette idée. Les femmes seniors se montrent même plus intéressées par des hommes plus jeunes que par des hommes plus âgés. Ainsi, l'écart d'âge moyen entre les utilisatrices et leurs interlocuteurs passe de 6,6 ans en moyenne en faveur de l'homme chez les 18-24 ans, à 3,9 ans en faveur de la femme chez les 60-70 ans.

Chez les hommes, les tendances suivent une même tendance : l'intérêt va croissant pour un partenaire plus jeune avec l'âge. C'est ainsi que les utilisateurs de 18-24 ans échangent souvent avec des femmes aînées – l'écart d'âge avec leurs interlocutrices est de 2,2 ans en moyenne en faveur de la femme – alors qu'à l'autre bout de l'échelle des âges, les hommes sollicitent des femmes bien plus jeunes – 7 ans en moyenne.

Ces observations indiquent que l'écart d'âge est le résultat commun de processus différents. En début de parcours affectif, ce sont surtout les femmes qui réclament cet écart, alors qu'aux âges qui correspondent aux deuxièmes unions, ce sont plutôt les hommes qui revendiquent une différence d'âge en leur faveur. Alors que la force statistique de l'écart d'âge conduit souvent à y voir une *préférence forte et symétrique des deux sexes*, il est sollicité tantôt par les femmes, tantôt par les hommes.

Plus précisément, cette caractéristique courante des couples apparaît comme un « arbitrage » entre les préférences féminines et masculines qui divergent souvent par ailleurs. La figure 2 confronte les critères d'âge déclarés par les utilisateurs des deux sexes dans leur profil avec la différence d'âge effectivement observée dans les contacts établis. Elle montre que les rencontres se font à mi-chemin des préférences affichées par les femmes et les hommes. Plutôt qu'une *préférence partagée*, l'écart d'âge apparaît comme le compromis entre deux aspirations accommodées.

Ces observations remettent en cause la notion « choix » du conjoint, chère aux sociologues du couple. La rencontre amoureuse est une interaction complexe où les deux sexes s'affrontent et se frottent aux désirs de l'autre. Les données de *Meetic* révèlent les concessions et les renonciations qui font aussi les rencontres mais au sujet desquelles les enquêtes restent silencieuses.

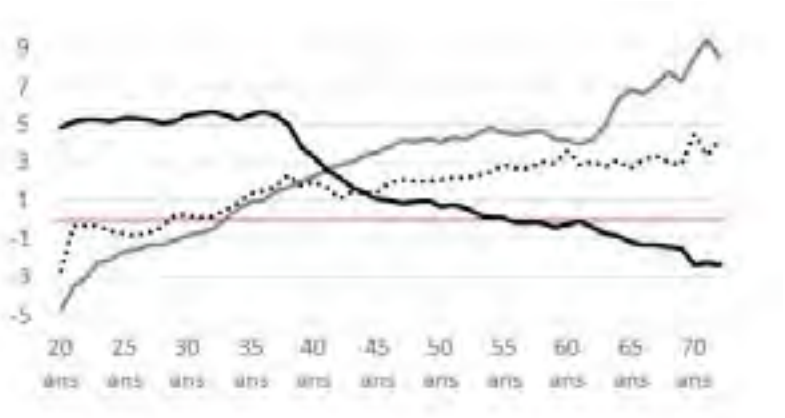


Figure 2. Écart d'âge moyen en faveur de l'homme tel que demandé dans les profils et observé dans les contacts établis[16]

16. **Lecture :** les profils présentant une femme de 55 ans affichent une préférence pour un écart d'âge moyen de 0,1 ans en faveur de la femme tandis que les profils présentant un homme de même âge affichent une préférence pour un écart moyen de 4,9 en faveur de l'homme. L'écart d'âge moyen des contacts établis par un expéditeur de 55 ans est de 2,7 ans.
Champ : compte d'utilisateurs actifs enregistrés sur *Meetic* en 2014 ; l'ensemble des premiers contacts ayant reçu une réponse sur *Meetic* en 2014.
Source : base d'utilisateurs de *Meetic.fr*, Meetic Group, 2014.

In fine, les sites de rencontres illustrent deux intérêts majeurs des données massives. D'une part, elles captent de façon avantageuses les *processus* qui, mal saisis par les enquêtes, sont pourtant centraux dans la réflexion en sciences sociales. D'autre part, elles offrent aux quantitativistes des possibilités nouvelles *d'observation* – une méthodologie jusque-là associée davantage au travail qualitatif. Ainsi, ces données ne sont pas seulement une source alternative mais proposent des *perspectives d'analyse statistique* nouvelles qui se soldent aussi par des résultats différents. Si les enquêtes conçues par les chercheurs restent inégalées de par leur richesse d'information, les données massives peuvent éclairer leurs zones d'ombres. Aujourd'hui, lorsque confrontés aux limites de leurs enquêtes, les chercheurs recourent aux hypothèses et aux « modèles d'acteurs ». La diffusion de nouvelles données nous incite à aborder autrement, et empiriquement, les problématiques à laquelle les questionnaires peinent à répondre. Cela implique de bricoler et d'accepter de travailler avec des données non représentatives. Les *Big Data* nous invitent en effet à élargir les critères de qualité des données quantitatives pour les juger autant au regard de leur pertinence et leur contenu informatif que de leur seul caractère représentatif.

SOCIÉTÉ

Le « Quantified self »

Marine BILLMANN, Valentine DELORME
Étudiantes
Ecole nationale de la statistique et de l'administration économique

Notre corps peut-il être un objet statistique, une réalité numérique ? La diffusion des objets connectés et des sites et applications permettant de conserver et interpréter des données sur le corps – tension, nombre de pas, température, ou même sur le comportement – humeur quotidienne, performances sexuelles – a-t-elle donné à leurs utilisateurs un corps statistiquement défini, un moi-quantifié ?

En 2007, quand Gary Wolf et Kevin Kelly, rédacteurs en chef du magazine *Wired*, lancent le « Quantified Self movement » (QS), ils parient sur l'émergence de telles identités, et encouragent leur développement. Pour eux, la collecte de données personnelles, leur interprétation et leur partage permettraient de progresser dans la connaissance de soi-même, de se réapproprier son corps et la gestion de son mode de vie : un programme ambitieux, qui questionne le rapport que notre société et ses individus entretiennent avec les statistiques et des données numériques. Le mouvement lancé par Wolf et Kelly se développe ensuite, autour du site internet *quantifiedself.com*, de rencontres internationales biannuelles, de « *meet-ups* » - i.e. de réunions, généralement lancées via le site meetup.com dans le contexte du QS - organisés régulièrement dans différentes grandes villes... D'après le site quantifiedself.com, en juillet 2015, le QS comptait 52000 membres dans le monde, réunis en 207 groupes, dans 37 pays. On peut noter que le mouvement a un écho certain dans le discours médiatique, et qu'il ouvre des perspectives académiques (notamment la création du *Quantified Self Institute*, qui cherche à étudier le vieillissement grâce aux données produites par la communauté) et médicales.

Le Quantified Self movement : vers un nouvel usage des statistiques individuelles ?

L'idée d'un « moi quantifié », d'une « auto-mesure de soi » - traductions courantes de *Quantified Self* - peut paraître assez intuitive : de la première pesée du nourrisson à la première course chronométrée de l'adolescent, des chiffres viennent caractériser le corps humain et ses performances depuis longtemps. Mais le « Quantified Self movement » postule l'apparition d'un nouveau moi quantifié : un moi défini par un plus grand ensemble de données, par des informations plus variées, et par sa capacité à faire progresser l'individu dans sa connaissance de soi et ses comportements.

Les données du Quantified Self sont d'un type nouveau. Les exemples que nous avons cités plus haut se rapportent à des mesures, des nombres. Or, pour reprendre la distinction faite

par Alain Desrosières[1], le Quantified Self ne consiste pas à *mesurer*, mais à *quantifier*. Dans la quantification, la production de données se fait en continu, immédiatement, et touche des comportements et caractéristiques qui étaient auparavant « exprimé[s] par des mots et non par des nombres ». La quantification est préalable à la mesure, elle permet la mesure de quelque chose qui n'était auparavant pas mesurable. Le moi-quantifié n'est pas uniquement constitué de données sur des éléments déjà mesurables, comme la taille, le poids ou le temps, mais il englobe aussi des éléments plus qualitatifs, comme l'humeur, le sommeil, la santé – et potentiellement, l'ensemble des éléments qui participent de notre « moi qualifié », pour reprendre l'expression de Swan[2]. Gary Wolf évoque ainsi l'exemple d'un « quantifier » qui enregistrait des données relatives à sa santé, mais également le nombre de films qu'il regardait, les personnes à qui il parlait, les sujets de conversations… Si les données produites sont donc plus variées, elles sont en outre récoltées beaucoup plus fréquemment, grâce à l'avènement des smartphones et autres objets connectés (de la montre à la brosse à dents en passant par le body) et à la progression des technologies de stockage et d'interprétation des données, qui permettent des collectes de mesure quasi-continues et peu contraignantes pour les utilisateurs.

Ce vaste ensemble de données a pour vocation d'être analysé par les « self-quantifiers », afin de gagner en connaissance d'eux-mêmes et d'améliorer leur comportement. Le Quantified Self a un objectif explicite de connaissance de soi, comme le souligne leur devise : « *Self knowledge through numbers* »[3]. Ce désir de connaissance induit un objectif d'amélioration de soi : la connaissance de soi n'est pas cherchée pour elle-même, mais en tant qu'elle permettra de modifier et d'améliorer ses pratiques. Pour certains membres du mouvement, l'amélioration des comportements repose sur des mesures relativement simples, sur un seul type de données qui permettra de s'astreindre à une discipline (par exemple, un écrivain pourra mesurer le nombre de pages qu'il écrit et se contraindre à en rédiger un certain nombre chaque jour), ou de mesurer ses performances (la vitesse atteinte lors d'une course). D'autres « self-quantifiers » ont un usage plus complexe des données, qui repose sur un postulat initial : la collecte d'un grand nombre d'informations leur permettra d'identifier des corrélations entre divers aspects de leur vie quotidienne et de chercher des facteurs explicatifs à certains de leurs comportements. Par exemple, des personnes souffrant de cyclothymie peuvent essayer de mettre en lumière les facteurs modulant leur humeur, que cela soit les interactions sociales qu'ils connaissent, les produits qu'ils consomment, leur quantité de sommeil… Si les données sont utilisées de manière plus ou moins complexe, elles répondent donc toujours à un double objectif : la connaissance et l'amélioration de soi.

La dernière nouveauté du « Quantified Self movement », c'est sa dimension collective : si la description des objectifs et des pratiques peut paraître très individualiste, les « self-quantifiers » se définissent comme un « mouvement », une « communauté ». Les données personnelles ne sont pas destinées à rester privées, mais doivent être diffusées, partagées au sein de la communauté du Quantified Self, sur les réseaux sociaux, les sites internet et les blogs, ou bien lors des « meet-ups » organisés par les différents groupes. Les membres du mouvement sont invités à expliquer ce qu'ils ont mesuré, comment ils l'ont mesuré, et ce qu'ils ont appris grâce à ces mesures. Les membres du groupe pourront donc comparer leurs expériences et leurs résultats, confronter leur moi-quantifié.

Une pratique émancipatrice ?

Le discours des fondateurs du Quantified Self est animé par l'idée que la quantification de soi permet une forme d'*empowerment*, mot que l'on peut traduire par « *mise en capacité d'agir* »,

1. Desrosières A. (2008), Pour une sociologie historique de la quantification, in. *L'argument statistique* I, Paris, Presses de l'École des Mines,
2. Swan M., (2013), The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery, Big Data, 1(2): 85-99
3. « La connaissance de soi par les nombres ».

grâce à la maîtrise de la production de ses données. Dans une société où nos traces numériques sont enregistrées sans que ne nous puissions vraiment y consentir, devenir producteur et analyste de ses propres données constituerait une forme de résistance contre une collecte de données subie, destinée à nous faire consommer, voire à nous surveiller, mais non à améliorer notre existence. De plus, le Quantified Self étend aux utilisateurs la possibilité de distinguer entre corps vécu et corps soigné et mesuré. En donnant accès à une connaissance experte de son organisme, le Quantified Self favoriserait le basculement d'une approche intuitive du corps à une appréhension médiatisée par ses données objectives, réservée jusqu'alors aux médecins. Les rapports aux soignants, au milieu médical s'en trouvent modifiés : se prévalant de ce « savoir profane », qui englobe auto mesure et recherche d'informations pour les interpréter, les patients se montrent plus exigeants à l'égard de leurs médecins, et cherchent à jouer un rôle plus actif dans le diagnostic et le choix du traitement.

Toutefois, une vision critique des pratiques d'auto-mesure, souvent d'inspiration foucaldienne, remet en cause la capacité du Quantified Self à être une source d'empowerment. La juriste Antoinette Rouvroy[4] dénonce ainsi une « fétichisation des données », à l'origine d'une « normopathie », bien éloignée de la revendication de réappropriation de soi présentée comme fin du Quantified Self. Selon cette critique, les « self-quantifiers » intérioriseraient sans recul suffisant des normes sociales et de bien-être, ce qui les conduit à une recherche sans limite de performance. En effet, comme elle est fondée sur la comparaison interindividuelle, la norme est mouvante, et le chiffre produit potentiellement jamais assez « normal » ou bon. En altérant la capacité à se fonder sur la perception intuitive de son corps, susceptible de favoriser la distance critique à l'égard de normes qui ne peuvent s'appliquer de façon uniforme à tous, l'appréhension objective de soi par des applications à l'ergonomie étudiée, auxquelles les individus s'en remettraient aveuglément, serait plus aliénante qu'émancipatrice. En outre, les données produites par les applications et objets connectés peuvent être récupérées par leurs créateurs et fabricants à des fins marchandes, comme le souligne la sociologue Deborah Lupton[5]. L'essayiste Evgeny Morozov[6] pointe les affinités entre le Quantified Self et une société de type néolibéral, dans laquelle les individus, enjoints d'être entrepreneurs d'eux-mêmes, se voient déléguer la responsabilité de leur santé, au détriment de réformes structurelles pour traiter les problèmes de santé publique. Une traduction possible serait, par exemple, que la surveillance individuelle poussée de son poids permise par les objets connectés laisse les États se dispenser de mettre en œuvre des politiques de prévention de l'obésité. Cette individualisation de la santé ouvrirait la voie à un affaiblissement de l'État-Providence. En raison aussi bien de la récupération marchande des données produites par les « self-quantifiers », de l'injonction à la prise en charge de soi qui envisage les individus comme des consommateurs plus que comme des citoyens, ou des risques de surveillance associés à la production de données personnelles, le Quantified Self est dénoncé comme un mouvement porteur d'un individualisme où l'accent mis sur la performance, la responsabilité et l'amélioration de soi fragiliserait l'individu lui-même et potentiellement les solidarités collectives.

Ces critiques tendent à considérer les « self-quantifiers » comme les acteurs assez passifs et peu réflexifs d'un système technologique, et donnent une vision unificatrice du phénomène. Toutefois, la question de la réflexivité des « self-quantifiers » requiert un examen des pratiques et du sens qu'ils leur donnent, qui révèle aussi la diversité des usages de la quantification de soi. Cette analyse des pratiques et des discours des quantifiers est à l'origine de « critiques de la critique », qui relativisent, voire contestent l'idée de « fétichisation des données », de soumission

4. In. CNIL (2014), Le corps, nouvel objet connecté ? Du quantified-self à la M-santé : les nouveaux territoires de la mise en données du monde, *Cahiers IP n°2 Innovation et Prospective*
5. Lupton D. (2014), Self-tracking modes: reflexive self-monitoring and data practices, Paper for the 'Imminent Citizenships: Personhood and Identity Politics in the Informatic Age' workshop, 27 August 2014, ANU, Canberra
6. Cité in CNIL (2014)

aux normes sociales. L'anthropologue Dorien Zandbergen[7] et la philosophe Tamar Sharon , s'appuyant sur des analyses de terrain, rejettent l'idée de passivité des « self-quantifiers », et distinguent trois attitudes principales envers l'auto mesure : l'auto mesure comme « pleine conscience », qui se réfère à la forte présence à soi-même à laquelle le Quantified Self permet d'accéder, l'auto mesure comme « résistance », catégorie qui relaie le discours des fondateurs, et se nourrit d'exemples de patients qui ont pu, en dépit des réticences des médecins, devenir acteurs de leur traitement ou diagnostic, et l'auto mesure comme pratique narrative et communicative. Par la mise en récit de leurs pratiques de mesure et de leurs analyses, elles soutiennent que ceux qui adhèrent au Quantified Self, plus que ceux qui pratiquent l'auto mesure de façon occasionnelle ou dans un cadre strictement privé, acquièrent un réel recul sur leurs pratiques, et aboutissent autant à une qualification qu'à une quantification de leur moi.

D'autres analyses fondées sur une description des pratiques donnent une vision plus nuancée, plus attentive aux ambiguïtés de la mesure. Ainsi, Pharabod et ses collègues[8] mettent en avant la réflexivité des « self-quantifiers », qui mettent en question les données produites – ce qui peut passer par la recherche de moyens propres d'auto mesure et le rejet des applications – mais soulignent aussi les échecs possibles de la mesure, la difficulté à rendre intelligibles ses données. Ces difficultés peuvent détourner certains de la pratique, ou mener à une subjectivation des mesures, alors même que le chiffre est censé objectiver le corps, avec pour conséquence une focalisation sur certaines des mesures plus propices à l'interprétation, qui peut faire perdre une vision globale de soi.

Ambiguë dans sa capacité à émanciper l'individu, la pratique du Quantified Self l'est aussi dans sa dimension collective. Le partage des données personnelles des « self-quantifiers » est souvent dénoncé comme une manifestation d'égoïsme ou de volonté d'exposition de soi , et les autres impliqués dans son processus de mesure seraient avant tout les spectateurs d'une mise en scène de son identité. Les promoteurs du Quantified Self insistent au contraire sur ce que peuvent apporter à la compréhension de ses propres données la comparaison et la communication autour des résultats obtenus. Des études de terrain, il ressort que le sens des échanges de données personnelles, dont la collecte manque de systématicité, même parmi les « self-quantifiers » et dont l'appui normatif n'a pas nécessairement des contours bien définis, ainsi que la capacité du Quantified Self à dépasser la présentation d'un soi potentiellement calculé et à faire participer les autres adeptes du Quantified Self à son projet pour soi, sont incertains. Le caractère autocentré des pratiques initiales du mouvement a toutefois mené à une inflexion des discours : le magazine *Wired*[10] en a appelé au passage du Quantified Self au Quantified Us en 2014, défendant l'idée que la mesure de soi ne pourrait avoir d'utilité que si elle s'inscrit dans une communauté centrée autour de problématiques partagées, cherchant à donner sens collectivement à leurs données, et cite par exemple Crohnology, réseau social de patients souffrant de la maladie de Crohn. Le potentiel de contribution au bien commun du Quantified Self reste toutefois à mettre à l'épreuve des faits. De plus, ce type de projet collectif n'échappe a *fortiori* pas au risque de surveillance à grande échelle induit par le Quantified Self.

Enjeux juridiques du Quantified Self

Quoique les données personnelles soient protégées par la loi informatique et libertés de 1978, le développement des objets connectés a ouvert un nouvel univers technologique face auquel cette loi se révèle insuffisante[11]. La difficulté repose en particulier dans le caractère flou des

7. Sharon T.et, Zandbergen D. (2016), From data fetishism to quantifing selves: Self-tracking practices and the other values of data, *New Media & Society*
8. Pharabod A. et al. (2013), La mise en chiffres de soi. Une approche compréhensive des mesures personnelles, Réseaux 2013/1 (n° 177), 97-129.
9. Gicquel C.et Guyot P. (2015), *Quantified Self, les apprentis sorciers du « moi connecté* , Limoges, Éditions Fyp.
10. https://www.wired.com/2014/04/forget-the-quantified-self-we-need-to-build-the-quantified-us/
11. Lanna M. (2016), Le Quantified Self, nouveau moteur du big data et menace pour la vie privée, *Petites affiches* n°95.

données récoltées : elles ne sont pas répertoriées comme des données de santé, et donc pas protégées à ce titre, mais elles sont éminemment intimes. La Cnil conclut donc à la nécessité de leur reconnaître un statut particulier qui pourrait leur permettre de s'ancrer dans une régulation clairement définie.

La loi pour une République numérique du 7 octobre 2016, qui a pour objectif de permettre aux individus de « mieux protéger leurs données personnelles », répond en partie aux défis posés par le Quantified Self. Elle affirme le droit à l'autodétermination informationnelle, le droit à l'oubli pour les personnes mineures, et la possibilité de demander à ce que les données personnelles soient effacées après la mort. Les individus doivent également être informés de la collecte de leurs données, de la durée de conservation de celles-ci. Les pouvoirs de la Cnil sont également renforcés pour pouvoir sanctionner les organismes enfreignant la loi. Cette loi se place dans la lignée du principe juridique du « *Privacy by design* »[12], qui demande à ce que la protection de la vie privée soit intégrée dans le développement de nouveaux services ou applications (par exemple, en indiquant clairement quelles données seront collectées et quel sera leur usage quand un utilisateur installe une nouvelle application). Cependant, certains juristes considèrent que ce principe n'est pas suffisant, et qu'il doit être remplacé, ou complété, par celui de « *Privacy by Using* », qui vise à réduire l'asymétrie d'information entre l'usager et le créateur en incitant les individus à protéger leur vie privée et à avoir une utilisation éclairée des applications et objets connectés[13].

Conclusion

Le Quantified Self est un mouvement qui pourrait paraître anecdotique au regard du nombre de personnes se définissant comme « self-quantifiers », mais dont la dynamique dépasse les frontières du mouvement lancé par Wolf et Kelly. Au-delà des membres du mouvement, ou même de ceux qui partagent leurs données, des pratiques privées comme le dénombrement de ses pas au moyen d'objets comme les montres connectées témoigne de la diffusion de la pratique de l'auto mesure, solidaire de nouveaux rapports au corps et de l'importance donnée à la performance.

Ainsi, le développement du Quantified Self, s'appuyant sur une nouvelle perception des données et une redéfinition du domaine de la quantification, soulève de nombreuses questions en termes d'autonomie, de responsabilisation et de protection des individus. Malgré sa faible ampleur, le Quantified Self pourrait être le signe d'une transformation globale des modes de vie et de l'économie, notamment dans le domaine de la santé, et rend plus aiguë la nécessité de réfléchir aux implications politiques de l'extension de la quantification et de la place du numérique dans la société.

12. NDR : « confidentialité assurée dès la conception »
13. Rallet A., Rochelandet F. et Zolynski C.(2015), De la Privacy by Design à la Privacy by Using. Regards croisés droit/économie, *Réseaux*, 2015/1 (n° 189), 15-46.