Un air de famille : les trajectoires parallèles de l'open data et du big data

Samuel Goeta

27 800 signes (hors bio et résumé)

Un air de famille : les trajectoires parallèles de l'*open data* et du *big data*

Samuel Goëta – sociologue

(chapô 429 s.)
Il n'est pas rare de voir les commentateurs confondre *open data* et *big data*, deux termes apparus vers 2008 pour désigner des pratiques émergentes de diffusion et d'exploitation des données numériques. Mais les données ouvertes ont rarement les attributs du *big data*, à commencer par le volume. De plus, les données ouvertes sont exploitables librement et gratuitement alors que les données massives restent un trésor bien gardé.

(Article)
Dans le journal *Le Monde* du 11 mars 2015, la deuxième page est consacrée à un portrait d'Henri Verdier, administrateur général des données et directeur d'Etalab, la mission en charge de l'ouverture des données publiques. L'article traite uniquement de l'ouverture des données (*open data*) de l'État alors qu'il est publié à l'occasion du salon Big Data qui se tient à Paris à ce moment-là. La confusion opérée par cet article du *Monde* n'est pas un cas isolé. Très souvent, le discours commun ne fait pas la distinction entre *open data* et *big data*, deux notions à la sonorité proche qui réapparaissent régulièrement dans les débats sur les technologies numériques. Tous deux popularisés à partir de 2008, l'*open data* et le *big data* ont en commun de susciter l'espoir d'une nouvelle manne économique et l'apparition de connaissances jusqu'alors inenvisageables. Toutefois, confondre ces deux termes, c'est ignorer les différences radicales de leur mise en œuvre, en particulier leurs pratiques de diffusion et d'exploitation des données. L'*open data* désigne la diffusion volontaire et proactive de données, essentiellement publiques, qui deviennent librement réutilisables alors que le *big data* regroupe des pratiques d'exploitation de données massives, à grande vitesse et en temps réel.
Cet article explicite dans un premier temps les origines de l'*open data* et du *big data* et rend compte des différences entre les motivations et les pratiques à l'œuvre dans ces deux champs. Alors que la transparence, l'innovation ou la modernisation de l'État servent de justification à l'*open data*, un impératif d'efficience et de prédictibilité fondent les pratiques du big data. Les données ouvertes se distinguent aussi par leur régime juridique, qui les rend librement réutilisables par tous. À l'inverse, les données massives restent la plupart du temps protégées et sont rarement réutilisables en dehors des frontières des organisations.

Open data : une injonction à la transparence et l'innovation

L'*open data* trouve son origine dans deux mouvements principaux : le droit d'accès aux informations du secteur public et l'ouverture des données scientifiques. Il s'inscrit dans l'histoire longue de l'accès du public aux informations produites par l'administration. Déjà, en 1789, la Déclaration des droits de l'homme et du citoyen stipule dans son article 15 que « *la Société a le*

droit de demander compte à tout Agent public de son administration » faisant de l'accès à l'information publique un des fondements de la démocratie naissante. Une autre racine des politiques d'*open data* est la notion d'*Open Government* apparue aux États-Unis à la suite de la Seconde Guerre mondiale et invoquée pendant la Guerre du Vietnam pour exiger de l'État fédéral qu'il réduise son opacité (Yu et Robinson, 2012). En 1966, le *Freedom of Information Act* donne ainsi le droit à chaque citoyen américain d'exiger de l'administration les informations qu'elle détient. En France, la loi Cada, du nom de la Commission d'accès aux documents administratifs, établit en 1978 un droit similaire qui permet à chaque citoyen d'exiger à une administration la publication de ses informations publiques. Cette loi définit celles-ci comme les informations produites dans le cadre d'une mission de service public (Boustany, 2013). Cette définition légale exclut les données personnelles du champ des politiques d'*open data* en France et dans la plupart des pays européens. C'est là une différence majeure avec les techniques de *big data* qui ont souvent recours à de volumes conséquents de données personnelles et nominatives. Aujourd'hui, ce droit d'accès aux données publiques est reconnu comme une liberté fondamentale dans la plupart des démocraties. Les politiques d'*open data* le prolongent en incitant les administrations à diffuser leurs données sans attendre qu'un citoyen exerce son droit d'accès tout en garantissant leur anonymat. En outre, dans le champ scientifique, par exemple dans des disciplines comme la botanique, la génétique ou l'astronomie, les pratiques de partage de données sont devenues monnaie courante (Bowker et al, 2010 ; Edwards *et al.*, 2011 ; Strasser, 2012). Le partage des données permet de réduire les coûts des recherches, de développer les pratiques comparatives, de rendre plus transparente l'exploitation des données ou de faciliter la vérification et la reproduction des protocoles scientifiques (Borgman, 2007). C'est d'ailleurs dans un rapport de l'Académie des sciences américaine suggérant le partage libre des données environnementales collectées par satellite qu'est apparu le terme « *open data* » en 1995 (1).

Alors que les pratiques d'ouverture de données sont déjà bien établies dans le monde académique, l'*open data* en tant que revendication politique fait son apparition au Royaume-Uni en 2006 à la suite d'une tribune dans le quotidien *The Guardian* (2). Son argument, devenu un des fondements de l'*open data*, consiste à revendiquer l'ouverture de ces données au motif qu'elles sont produites avec l'argent des contribuables. Les principes de l'*open data* ont été posés lors d'une rencontre qui s'est tenue à Sebastopol en Californie en 2007, organisée par des acteurs importants du numérique tels que l'éditeur Tim O'Reilly, connu pour avoir forgé l'expression « web 2.0 », l'avocat Lawrence Lessig, à l'origine des licences Creative Commons qui proposent une alternative au *copyright* en permettant le partage libre et gratuit des œuvres culturelles, ou Aaron Swartz, militant de l'ouverture du savoir connu entre autres pour avoir cofondé le site communautaire de partage de liens Reddit. Leur influence a entraîné une diffusion rapide des « principes de Sebastopol » (3), lesquels exigent la publication des données dès leur production et dans leur forme « primaire », c'est-à-dire telles que les administrations les collectent, sans avoir été modifiées ou agrégées. Le président Obama s'en est inspiré pour signer, le jour de son entrée à la Maison-Blanche, un mémorandum sur l'*Open Government* qui a abouti à la création de data.gov, premier portail de diffusion de données gouvernementales ouvertes. Il s'en est suivi, par mimétisme, une prolifération de portails diffusant des données publiques aux niveaux local comme national, tels data.gov.uk en 2009 au Royaume-Uni et data.gouv.fr en 2011 en France. En 2013, les principes de l'*open data* ont

été repris par le G8 qui établit dans une charte que l'ouverture des données est la pratique par défaut des administrations des huit pays signataires.

L'*open data* a ainsi pris la forme d'une injonction, d'une demande d'ouverture des données publiques portée par une multitude de groupes d'intérêt. Parmi ces derniers, on peut citer des associations (l'Open Knowledge Foundation ou la Sunlight Foundation au niveau mondial, Regards Citoyens et Libertic en France), des *think tanks* (la Fondation Internet nouvelle génération (Fing), la Fondation pour l'innovation publique ou Terra Nova) ou encore des acteurs publics (comme Étalab qui a pour mission de promouvoir l'*open data*). Tous ces acteurs ont justifié l'ouverture des données par ce que Dodier (2003) qualifie de « biens en soi », des objectifs digne d'être poursuivis en tant que tel sans nécessiter de justification supplémentaire. Trois notions principales peuvent résumer ces « biens en soi », même si plusieurs autres peuvent être identifiées (Denis & Goëta, 2015) : la transparence, l'innovation et la modernisation de l'État.

L'ouverture des données est souvent considérée comme une forme supérieure de transparence, réputée plus objective que les processus de révélation qui l'ont précédée (Birchall, 2014). La publication de données brutes, estimées objectives et non interprétées, permettrait d'examiner le fonctionnement de l'État au niveau le plus local, en une forme de surveillance citoyenne considérée comme un renouveau voire un paroxysme de la transparence (Goëta, 2015).

Concernant l'innovation, les politiques d'*open data* ont été aussi conçues comme une incitation au développement de l'industrie de l'information. Dans les années 2000, la Commission européenne a promu l'ouverture des données comme un nouveau filon économique à exploiter. Elle a évalué la valeur économique de la diffusion gratuite des données publiques à plusieurs dizaines de milliards d'euros par an dans l'Union (Vickery, 2011). En 2003, elle a adopté la directive Public Sector Information (PSI) pour inciter les États membres à diffuser gratuitement leurs données. D'autre part, les politiques d'*open data* ont été promues dans le contexte de développement de l'économie des applications mobiles à partir de l'année 2007. Il est attendu que les données ouvertes soient réutilisées pour créer des services pratiques sur mobile.

Enfin, les politiques d'ouverture des données sont justifiées par leur capacité à provoquer des changements dans le fonctionnement de l'administration et à engendrer ainsi sa « modernisation. » Les services dédiées à la transformation de l'administration, comme le secrétariat général pour la modernisation de l'action publique (SGMAP), auquel Etalab est rattaché, soutiennent fortement les politiques d'*open data*. L'ouverture des données décloisonnerait les services et favoriserait le travail au-delà des divisions administratives. Entre réforme de l'État et nouvelles opportunités de communication, l'ouverture des données publiques s'inscrit dans la lignée des projets d'administration électronique mis en œuvre depuis les années 1990 (Dagiral, 2011).

Big Data **: la promesse d'un renouvellement de la connaissance et d'une plus grande efficacité des organisations**

La notion de *big data* a émergé à la même période que celle d'*open data*, entre 2007 et 2008. Le terme et les pratiques qu'il recouvre ont également des origines bien plus anciennes, dans une moindre mesure que pour l'*open data*. L'entreprise Silicon Graphics Inc. (SGI) avait employé l'expression *big data* dès le milieu des années 1990 pour désigner ses solutions d'exploitation de données massives. Les premières références académiques au *big data* apparaissent quant à elles

dès 1998 (Diebold, 2012). L'année 2008 a marqué un tournant : d'une part, la presse s'est saisie du terme, notamment le magazine *Wired* et la revue scientifique *Nature* qui ont chacun fait leur une sur la science à l'« âge du pétabit » (4) ; d'autre part, un consortium regroupant des acteurs importants de la recherche en informatique aux États-Unis, dont la National Science Foundation, a publié un rapport qui a donné une crédibilité académique à la notion de *big data* (5). Après la publication de ce rapport, de grandes entreprises telles que IBM ou SAS ont fait du big data un slogan pour vendre leurs solutions informatiques (Lohr, 2012). Aujourd'hui, le *big data* est communément défini par trois caractéristiques, largement reprises par la littérature (Zikopoulos *et al.*, 2012 ; Kitchin, 2014), et connues comme les trois « V » du *big data* : le *volume*, avec l'exploitation de térabits voire de pétabits de données, la *vélocité* de leur traitement quasiment en temps réel et la *variété* des sources traitées. Kitchin (2014) complète cette définition par plusieurs autres caractéristiques : l'*exhaustivité* des données, avec l'analyse de populations entières plutôt que d'échantillons, la *résolution* des données, qui se situe au niveau le plus local sans agrégation, et la « *scalability* » (traduite par « l'extensibilité » ou « l'évolutivité ») des données, c'est-à-dire la mise en *relation* de sources de données très diverses et la capacité de ces systèmes à croître rapidement. Enfin, il faut souligner l'importance de nouveaux systèmes de bases de données caractéristiques du *big data* (6) développés à partir de 2007 pour traiter de grands volumes de données. Contrairement aux systèmes de bases de données relationnelles utilisés dans les systèmes d'information depuis les années 1970, ils n'exigent pas que soient déterminées à l'avance les relations entre les éléments contenus dans la base (Driscoll, 2012).

Le *big data* a été justifié globalement parce qu'il induit une plus grande efficience du fonctionnement des organisations. Dans les administrations publiques, il a trouvé une première application dans la gestion des villes par le traitement en temps réel des données collectées par des capteurs omniprésents dans l'espace urbain. Des algorithmes ont exploité les données pour prédire les zones de criminalité et y faire intervenir les forces de police préventivement (Harcourt, 2007). Dans les entreprises, les techniques de *big data* ont été promues pour accroître la productivité des équipes, prédire l'évolution des prix ou encore optimiser l'allocation des ressources (Kitchin, 2014). En particulier, les services de marketing ont recours à l'exploitation de données massives pour cibler leurs actions auprès de segments de clientèle très fins (Manyika *et al.*, 2011). Dans les sciences, le *big data* a été présenté comme une révolution scientifique pouvant profondément renouveler les pratiques de production du savoir. Par exemple, Chris Anderson, rédacteur en chef du magazine *Wired*, annonçait en 2009 la « fin de la théorie ». Selon cet auteur, connu pour ses prophéties technologiques, les volumes de données à analyser sont tels que les chercheurs peuvent partir des corrélations pour formuler des hypothèses plutôt que de s'appuyer sur la littérature et la théorie. À la suite, des chercheurs de Microsoft ont annoncé en 2009 un « quatrième paradigme », dans lequel il suffirait de laisser parler les données pour comprendre les phénomènes (Hey *et al.*, 2014), le risque étant de créer de nouvelles fractures entre les chercheurs « riches en données » et ceux n'y ayant pas accès (Manovich 2011 ; Boyd & Crawford 2012}. Très critiqué, ce nouveau positivisme considère les données comme naturellement objectives sans prendre en compte le contexte théorique et matériel de leur production (Bowker, 2000; Kitchin, 2013 ; Ribes, 2013).

En résumé, dans les administrations, les entreprises et les sciences, le *big data* a été associé à la

promesse d'une plus grande efficience et de nouvelles formes de savoir caractérisées par la prédiction.

Les données ouvertes : des ressources librement utilisables

Contrairement à l'*open data*, la plupart des systèmes de *big data* ne partagent pas les données qu'ils utilisent. Elles restent à l'intérieur des systèmes d'information de l'organisation et les algorithmes qui les traitent constituent des boîtes noires dont le fonctionnement est généralement méconnu du public. Les données sont parfois diffusées par des interfaces de programmation (*API* ou *Application Programming Interface*) qui en définissent les conditions d'accès et en restreignent les usages. Facebook, Google ou Twitter fournissent de telles interfaces qui permettent aux usagers de partager une sélection de leurs données avec une application tierce. Toutefois, les interfaces de programmation ne suffisent pas à ouvrir les données au sens des principes fondateurs de l'*open data.* Elles ne donnent accès qu'à un échantillon des données dans des conditions très restrictives, qui restent généralement la propriété de l'entreprise qui les a collectées.

À l'inverse, l'*open data* se caractérise par un régime juridique qui incite à la réutilisation des données. L'Open Knowledge Foundation a défini, à partir de 2005, les conditions de l'ouverture d'une donnée à travers l'*Open Definition*, laquelle fait aujourd'hui référence pour délimiter ce que désigne l'appellation « donnée ouverte ». Cette définition stipule qu'une œuvre (et donc une donnée) est ouverte si quiconque est libre de l'utiliser, la réutiliser et la redistribuer. L'*Open Definition* autorise deux conditions : citer la source et/ou demander le partage avec la même licence des données modifiées. En plus de ce régime juridique qui encourage à la réutilisation, les données ouvertes doivent être accessibles dans leur intégralité. La simple mise à disposition d'une interface de programmation ne suffit pas à ouvrir des données ; ces dernières doivent être librement réutilisables sans inscription préalable. Ces conditions sont rarement remplies par les bases de données du *big data* dont les usages sont généralement restreints à la fois par leur caractère stratégique et par la présence de données personnelles devant être anonymisées.

Les cas de bases de données à la fois massives et ouvertes restent rares. Les données publiées par les gouvernements répondent rarement aux caractéristiques du *big data* évoquées précédemment. Dans la plupart des cas, il s'agit de ce que Kitchin (2014) qualifie de « *small data* », des données de faible volume qui agrègent une base plus massive, en fournissent un extrait ou sont bâties sur un échantillon de population. Ainsi, la majorité des données publiées sur les portails d'*open data* ne dépassent pas le mégaoctet. Leur mise à jour est essentiellement périodique, annuelle ou mensuelle, exceptionnellement quotidienne ou en temps réel. Le faible nombre de données massives et ouvertes du secteur public s'explique en grande partie par la difficulté de garantir l'anonymisation de bases de données contenant des informations très sensibles sur les individus (6). D'autre part, les systèmes d'information de l'État dépendent de logiciels propriétaires qui verrouillent l'accès aux données. Créer des passerelles entre ces derniers et les portails *open data* requiert une opération complexe d'exploration des infrastructures de production et d'exploitation des données (Denis et Goëta, 2014).

Dans le secteur public, la surveillance de l'environnement fournit les données les plus massives, mais leur ouverture n'est pas systématique du fait notamment des redevances encore en place (Trojette, 2013). En Europe, seul l'institut météorologique finlandais fournit l'intégralité de ses

données en *open data*. En France, les bases de données dans le secteur de la santé qui pourraient être ouvertes sont nombreuses (7), mais leur ouverture reconfigurerait le système de santé en donnant aux services de l'État un accès direct aux bases de données des caisses d'Assurance maladie (Briatte & Goëta, 2014). La plupart des demandes se concentrent sur le Système national inter régimes d'Assurance maladie (SNIIRAM) dont une première ouverture vient d'être réalisée avec une extraction mensuelle des dépenses d'assurance maladie (voir l'article d'Hélène Caillol). Le volume de ces données détaillées est considérable, une extraction de la base dans un seul fichier statique dépasserait le milliard de lignes (8). Les données publiques sont aussi traitées et agrégées par des acteurs privés et associatifs pour constituer des bases massives et partagées. En France, l'association Regards Citoyens extrait automatiquement les données des sites de l'Assemblée nationale et du Sénat afin de développer des outils de surveillance de l'activité parlementaire. De manière similaire, le site Open Corporates regroupe les données des registres des entreprises de 75 juridictions dans le monde qui sont exploitables même pour des usages commerciaux.

D'autres bases de données massives et ouvertes sont collectées grâce au travail d'une multitude de contributeurs. Ces ressources sont un bien commun, partagé et administré par ses contributeurs. Le cas le plus connu est celui d'OpenStreetMap (OSM), une base de données géographique mondiale qui repose sur les contributions des citoyens qui éditent le « Wikipedia de la carte ». OSM couvre désormais la plupart des pays du monde et concurrence le géant Google Maps. Dans les sciences, TeleBotanica regroupe des botanistes professionnels et amateurs, qui créent collaborativement une base de données partagée sur la faune et la flore partout dans le monde. On peut aussi citer OpenFoodFacts pour les données nutritionnelles, OpenMeteoData pour la météorologie ou encore le réseau Citoyens Capteurs en matière de pollution.

***

La proximité sémantique entre *big data* et *open data* prête à confusion, tout comme leur émergence simultanée. Les deux mouvements ont suscité des espoirs en matière de création de savoir et d'efficience des organisations mais aussi des craintes portant essentiellement sur les atteintes à la vie privée qui pourraient survenir lors de l'exploitation des données. Néanmoins, *open data* et *big data* divergent par leurs finalités et leur régime de partage des données. Ainsi, les cas de données à la fois ouvertes et massives restent encore rares. L'ouverture des données massives du secteur public implique une transformation des systèmes d'information de l'État, une réflexion sur les modèles économiques de l'ouverture et un approfondissement des travaux sur l'anonymisation des données.

Notes

1 – National Academy of Sciences (1995). « On the Full and Open Exchange of Scientific Data », *http://www.nap.edu/readingroom.php?book=exch&page=summary.html*

2 – Arthur, C., & Cross, M. (2006, 9 mars), « Give us back our crown jewels », The Guardian, *http://www.theguardian.com/technology/2006/mar/09/education.epublic*

3 – 8 Principles of Open Government Data, *https://public.resource.org/8_principles.html*

4 – Le volume des disques durs de nos ordinateurs personnels dépasse rarement le térabit, soit un millième de pétabit.

5 – Computing Community Consortium (2008). « "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society », http://www.cra.org/ccc/docs/init/Big_Data.pdf.

6 – « NoSQL » (*Not only SQL*) a été proposé en 2009 pour désigner les systèmes de gestion de bases de données employés par les géants du web tels que Facebook, Google ou Amazon. Ces systèmes redéfinissent l'architecture classique des bases de données relationnelles et ne requièrent plus nécessairement l'emploi du langage SQL. Le *Structured Query Language* (SQL), « langage de requête structurée », est un langage informatique normalisé servant à exploiter des bases de données relationnelles.

7 – L'effacement des informations nominatives ne suffit pas à anonymiser complètement une base de données. Les identifiants uniques de chaque usager peuvent servir à ré-identifier les individus dans une base de données ne comportant aucun nom. Par exemple, à Londres, des données ont été ouvertes sur les déplacements des habitants en vélo à libre service (voir un article du site *Quartz* (2014) consulté à l'adresse *http://qz.com/199209/londons-bike-share-program-unwittingly-revealed-its-cyclists-movements-for-the-world-to-see*). La base de données comportait des identifiants numériques uniques pour chaque usager. En connaissant l'heure, le départ et la destination d'un individu lors d'au moins deux trajets, il était alors possible de retracer l'ensemble de ses déplacements. Pour garantir l'anonymat, il est donc souvent suggéré d'introduire des identifiants aléatoires pour empêcher de tels cas de ré-identification. En France, en 2014, un rapport du Sénat (rapport d'information n° 469) a signalé essentiellement un seul cas de publication de données pouvant donner lieu à réidentification concernant les données socio-économiques dites carroyées (dans des carreaux de 200m de côté) dans les zones de faible densité (l'INSEE a depuis changé sa méthodologie).

7 – Étalab a réalisé une cartographie des bases de données de santé disponibles en France. Voir Étalab, 2014, « Open Data en Santé : Publication de la cartographie des données publiques en santé et ouverture d'une consultation publique », *https://www.etalab.gouv.fr/opendataensantepublicationdelacartographiedesdonneespubliquesensanteetouvertureduneconsultationpublique*

8 – Étalab, 2015, Retour sur le premier Hackathon « données de santé », *https://www.etalab.gouv.fr/retour-sur-le-premier-hackathon-donnees-de-sante*

Bibliographie

Birchall C., 2014, « **Radical Transparency?** », *Cultural Studies - Critical Methodologies*, n° 14, p. 77-88.

Borgman C. L., 2007, ***Scholarship in the Digital Age. Information, Infrastructure and the Internet***, Cambridge, The MIT Press.

Boustany J., 2013, « **Accès et réutilisation des données publiques. État des lieux en France** », *Les cahiers du numérique,* n° 9, p. 21-37.

Bowker G. C., 2000, « **Biodiversity datadiversity** », *Social Studies of Science*, n° 643.

Bowker G. C. *et al.*, 2010, « **Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment** », *International Handbook of Internet Research,* p. 97-117.

Boyd D. et Crawford K., 2012, « **Critical questions for big data** », *Information, Communication*

& Society, vol. 5, n° 15, p. 662-679.

Briatte F. et Goëta S., 2014, « **Les logiques politiques de l'ouverture des données de santé** », *Statistique et Société,* n° 2, p. 49-55.

Dagiral É, 2011, « **Administration électronique** », *Communications*, n° 88, p. 9.

Denis J. et Goëta S., 2015, « **La fabrique des données brutes. Le travail en coulisses de l'open data** », *in Penser l'écosystème des données*, (à paraître), Paris, Éditions de la Fondation Maison des Sciences de l'Homme, *https://halshs.archives-ouvertes.fr/halshs-00990771*

Denis J. et Goëta S., 2015, « **Les facettes de l'open data : émergence, fondements et travail en coulisses** », *in Big data, entreprises et sciences sociales*, (à paraitre), Paris, Collège de France.

Diebold F., 2013, « **A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline**, *PIER Working Paper*, n° 13-003, Social Science Research Network.

Dodier N., 2003, ***Leçons politiques de l'épidémie de sida,*** Paris, Editions de l'EHESS.

Driscoll K., 2012, « **From Punched Cards to "Big Data": A Social History of Database Populism** », *communication +1*.

Edwards P. *et al.*, 2011, « **Science friction: Data, metadata, and collaboration** », *Social Studies of Science*, n° 41/5, p. 667-690.

Hey *et al.*, 2009, ***The Fourth Paradigm. Data-intensive scientific discovery,*** Microsoft Research.

Goëta S., 2015, « **L'open data : une forme ultime de transparence ?** », *in* Catellani A. *et al.* (dir.), *La communication transparente. L'impératif de la transparence dans le discours des organisations,* Louvain-la-Neuve, Presses Universitaires de Louvain.

Harcourt B. E., 2007, ***Against prediction. Profiling, policing, and punishing in an actuarial age,*** Chicago, University of Chicago Press.

Kitchin R., 2013. « **Big data and human geography: Opportunities, challenges and risks** », *Dialogues in Human Geography*, vol. 3, n° 3, p. 262-267.

Kitchin R., 2014, ***The Data Revolution. Big Data, Open Data, Data Infrastructures & their consequences***, Sage Publications.

Lohr S., 2012, « **How Big Data Became so Big** », *The New York Times,* 12 août 2012.

Manyika J. et al, 2011, ***Big data: The next frontier for innovation, competition, and productivity,*** San Francisco, McKinsey Global Institute.

Ribes, D. et Jackson S. J., 2013, « **Data bite man: The work of sustaining a long-term study** », in L. Gitelman L. (dir.), *Raw data is an oxymoron*, Cambridge, The MIT Press, p. 147-166.

Strasser B. J., 2012, « **Data-driven sciences: From wonder cabinets to electronic databases** », *Studies in history and philosophy of biological and biomedical sciences*, n° 43, p. 85-7.

Trojette. M., 2013, ***Ouverture des données publiques. Les exceptions au principe de gratuité sont-elles toutes légitimes ?***, rapport au Premier Ministre.

Vickery G., 2011, ***Review of recent studies on PSI re-use and related market***, Paris, éditeur.

Yu H. et Robinson D. G, 2012, « **The New Ambiguity of "Open Government"** », *UCLA Law Review*, n° 178, p. 178-208.

Zikopoulos P. *et al.*, 2012, ***Understanding Big Data. Analytics for Entreprise Class Hadoop and Streaming Data***, McGraw Hill.

Goëta Samuel

Doctorant au sein du département de sciences humaines de Telecom ParisTech. Ses recherches traitent principalement des enjeux politiques des technologies numériques. Sa thèse porte sur les coulisses de l'open data pour révéler comment les données sont sélectionnées, extraites et retravaillées avant leur publication. Il est aussi confondateur du chapitre français de l'Open Knowledge Foundation, une association qui œuvre pour que l'ouverture du savoir et des données bénéficie à tous.

Résumé

Le discours commun tend à confondre open data et big data, deux termes apparus simultanément aux alentours de 2008. Pourtant, les pratiques de diffusion et d'exploitation des données qui en résultent diffèrent profondément. D'une part, l'open data se distingue du big data par les modalités de son émergence. Alors qu'un impératif de performance guide l'essor du big data, les politiques d'open data aspirent à accroitre la transparence, soutenir l'innovation ou encore à moderniser les administrations. D'autre part, les données massives ne sont pratiquement jamais exploitables librement alors que les données ouvertes sont utilisables intégralement et par tous. Cet article évoquera enfin quelques exemples de bases de données massives et ouvertes.