Du traitement des données à la création de valeur : comprendre les pratiques professionnelles des réutilisateurs des données ouvertes

Valentyna Dymytrova, Françoise Paquienséguy

Du traitement des données à la création de valeur : comprendre les pratiques professionnelles des réutilisateurs des données ouvertes

Valentyna Dymytrova[1], Françoise Paquienséguy[2]

[1]Université Lyon 3, France - valentyna.dymytrova-baiov@univ-lyon3.fr

[2]Sciences Po Lyon, France – francoise.paquienseguy@sciencespo-lyon.fr

Résumé

A partir d'une enquête de terrain menée en France en 2017, cet article identifie différentes formes de réutilisation des données ouvertes et analyse les chaînes de traitement sur lesquelles elles se fondent. En décryptant ces chaînes et les outils mobilisés par trois catégories de réutilisateurs professionnels (développeurs, data scientists et data journalists), les auteurs discutent leurs liens avec la chaîne de création de valeur. Les pratiques et les attentes professionnelles y sont abordées, en termes de plus-value générée par les données, de modèle économique (le courtage informationnel) mais aussi de prestations de services innovants. Cet article est partiellement issu des travaux de l'ANR OpenSensingCity.

Mots-clés : open data, chaîne de valeur, traitement des données, développeurs, data scientists, data journalists

Abstract

Based on a field survey conducted in France in 2017, this article identifies different forms of open data reuse and analyses the processing data chains on which they are relied. By analysing the chains and the tools used by three categories of professional reusers (developers, data scientists and data journalists), the authors discuss their links with the value creation chain. Professional practices and expectations are also discussed in terms of value generated by data, of economic model (informational brokerage) but also of innovative service creation. This article results partially of the research ANR OpenSensingCity.

Keywords: open data, data value chain, data processing, developers, data scientists, data journalists

---

1. Introduction

Le processus d'ouverture des données renvoie d'abord à une politique publique de partage des données issues des registres de la statistique administrative ou collectées par les organismes publics et ensuite à leur mise en ligne à travers des portails et des plateformes dédiés (Boustany, 2013). Comme le soulignent Noyer et Carmès, l'ouverture des données publiques questionne les modes de gouvernance, les modèles économiques et les façons de penser la politique et l'espace public dans deux registres structurants (Noyer, Carmes, 2013). Le premier concerne les conditions d'accès aux données, aux informations et aux connaissances, définies par le processus de l'ouverture. Le deuxième concerne la production et la circulation de connaissances et de services dans de nouvelles formes et conditions, configurées par la réutilisation des données ouvertes (*Ibid*).

Si les cadres législatif et administratif de l'OD ont été récemment redéfinis par plusieurs lois, notamment, les lois Macron, NOTRe, Valter et Lemaire, et par le Plan d'Action National 2015-2017, rendre les données ouvertes ne

suffit pas à générer leur réutilisation (Kitchin, 2014). Celle-ci constitue un processus complexe impliquant plusieurs communautés professionnelles aux divers objectifs stratégiques et économiques et agissant au sein de divers cadres éthiques (Dymytrova, Paquienséguy, 2017). Malgré un nombre important de jeux de données publiques rendues disponibles ces dernières années le nombre d'applications exploitant les données ouvertes reste assez limité, tout autant que celui des utilisateurs et celui des services qui n'atteignent pas des seuils de viabilité (Turky, Foulonneau, 2015).

Réutiliser les données consiste en fait « à sortir des données de leur contexte initial de production pour leur offrir un nouveau cadre d'interprétation et de traitement dans de « nouveaux contextes sociaux » » (Labelle, Le Corf, 2012). Dans une approche info-communicationnelle de l'OD, l'ouverture et la réutilisation des données publiques peuvent s'envisager, sous l'angle des industries culturelles dont les données ouvertes serait une sous-filière au sens d'« une organisation de la chaîne du système de production d'un produit et surtout d'un groupe de produits, et ce jusqu'à la consommation » (Bouquillon, Miège, Mœglin, 2013). Acteurs de cette sous-filière, porteurs du modèle du courtage informationnel (Paquienséguy, 2016), les réutilisateurs qu'ils soient professionnels ou amateurs, valorisent l'open data *via* des applications qu'ils créent en transformant les données en services à destination des usagers, des clients, des consommateurs, des citoyens, du territoire.

Ainsi, les portails métropolitains Open data peuvent-ils s'éclairer aux principes du courtage informationnel tel que proposé par Mœglin et mieux situer les différentes chaînes de création de valeur et de traitement des données sur la base des éléments qui le définissent (Mœglin, 2005). Tout d'abord, l'entremise qui en assure la fonction centrale, partagée, commune et incontournable, cette fonction est portée par différentes catégories de réutilisateurs. Puis, la valorisation de l'intermédiaire qui se fait sur des modes complémentaires, ces modalités de rémunération ou de monétarisation vont de la vente de mots-clés au paiement à l'acte en passant par une rémunération indirecte (en termes de notoriété par exemple). Ensuite, le courtage lui-même, il est partiellement assuré par des logiciels ou des routines œuvrant à l'indexation comme à la recherche des données ; ici enfin, le travail de l'intermédiaire est valorisé pour en soi, en dépassant la fonction centrale qu'il assure.

Finalement, inscrire l'analyse de la réutilisation des données dans la logique du courtage informationnel conduit à un autre regard sur la place et les enjeux des chaînes de traitement des données et sur les modalités de création de la valeur, posture qui pose question : « Quels enjeux sous-tend l'information publique autant dans son mode de production, de recueil, de modalités de diffusion que sur l'organisation des acteurs de la sphère publique [...] ? » [Bardou-Boisnier, Pailliart, 2012). En fait, deux enjeux majeurs se dessinent au fil de la structuration de l'action publique territoriale à propos des données.

Le premier est celui de la valeur produite par les données et leur exploitation, autrement dit les données comme source de création de valeur. Si à l'échelle industrielle ou entrepreneuriale la formulation est une évidence, elle l'est beaucoup moins à propos de l'action publique et des données partagées ou ouvertes que les métropoles produisent et agrègent depuis la Directive Inspire, la réforme territoriale et les lois qui s'en sont suivies, comme la loi NOTRe par exemple. Les efforts des métropoles à encourager et accompagner la réutilisation des données ouvertes et les choix concrets mis en œuvre, comme les portails open data, la licence de réutilisation des données d'intérêt général, les bonnes pratiques en termes de données de mobilité urbaines, montrent bien qu'elles ne s'inscrivent pas dans la logique économique des industriels et professionnels des big data. Ainsi à travers la promotion de la réutilisation visent-elles à favoriser la création dans deux directions : la création de valeur pour les réutilisateurs et la création de services à destination des urbains. Elles cherchent ainsi à se réorganiser autour

de la donnée métropolitaine et à se mettre en ordre de marche vers la smart city ou du moins, vers une gouvernance inspirée par les données, sur les questions de mobilité principalement.

Le deuxième est justement celui de la transformation de l'action publique (Courmont, 2015) et de l'innovation qu'elle engendre ou qu'elle vise. En effet, toutes les métropoles ne visent pas la transformation en smart city, certaines se veulent villes créatives et d'autres villes durables. Remarquons au passage que le GrandLyon cumule ces objectifs dans ses projets de territoire et les visées économiques qu'ils portent. L'innovation organisationnelle vise donc à construire l'écosystème nécessaire à la réutilisation ou au partage des données métropolitaines, et tout particulièrement à partir des partenaires de proximité déjà engagés dans la production ou la réutilisation des données publiques et/ou privées en lien avec le territoire. La constitution même des directions des systèmes d'information des métropoles autour de la donnée[1], des chief data officiers ou encore des responsables des données métropolitaines est un des témoins, comme la mise en place, assez généralisées d'urban labs ou de living labs. Ces derniers, comme le Tubà à Lyon, le laboratoire d'innovation urbaine à Montréal ou le CityLab de Boston se veulent le lieu métropolitain d'innovation tant dans les méthodes de travail utilisées (agiles, collaboratives, partenariales) que dans les services qui résultent des actions qu'ils conduisent ou soutiennent. Partagée, la donnée métropolitaine ouvre la porte à l'innovation sur la base du bien commun et de l'intérêt général (Lévesque, 2007).

Le contexte et l'approche exposés ci-dessus, nous amènent à instruire deux questions principales : Quelles sont les chaînes de traitement de données qui les transforment en services innovants ? Comment le modèle de la chaîne de valeur permet-il d'inscrire la chaîne de traitement des données ouvertes dans la logique de la création ou de l'innovation en faveur du territoire ?

Après une brève présentation de la méthodologie, notre analyse des chaînes de traitement de données propres aux trois types de réutilisateurs professionnels (développeur, data scientists et data journalists) dévoilera différentes formes de réutilisations et leurs liens avec le modèle de la chaîne de valeur.

La communication s'appuie sur l'ANR OpenSensingCity 14-CE24-0029. La diversité des profils, des finalités et des pratiques des réutilisateurs des données, couplées à un objet d'étude très contemporain et en rapide progression, réclamait une méthodologie qualitative, sur la base d'entretiens avec différents réutilisateurs professionnels de données à l'échelle nationale : Paris, Lyon, Toulouse, Strasbourg, Nantes, Grenoble et Brest. Au total, 27 entretiens semi-directifs ont été conduits de février à avril 2017. Ont été interrogés 7 développeurs ; 6 data scientists/analysts, 6 data journalists, 3 fournisseurs/éditeurs de portails et 5 personnes ressources : chargés de mission et chefs de projet OD métropolitains et les fondateurs de la coopérative Datactivi.st. L'ensemble des documents méthodologiques est disponible (Dymytrova, Larroche, Paquienséguy, 2018).

2. Open et Big data : Chaîne de traitement des données et création de valeur

Big et Open, ces termes coexistent et se côtoient sans cesse. Bien des acteurs et des politiques engagés témoignent que gigantisme et mise à disposition sont les éléments fondamentaux porteurs d'innovation par le fait même qu'il engendre la participation d'une part et la diffusion de l'autre. Alliés à la participation introduite par le web 2.0, ils relèvent d'idéologie différentes ce qui rend difficile l'existence des données ouvertes et des idéaux d'open-government (Paquienséguy, 2020, p. 2-5) qui les ont faites naître. En effet, aucun professionnel de la donnée interrogé ne travaille dans le contexte exclusif de la donnée ouverte.

[1] Renommées pour l'occasion Direction à l'innovation numérique et aux systèmes d'information (DINSI)

Les modalités et les conditions d'utilisation des données dépendent de l'utilisation de la donnée, des besoins des clients et des usages pressentis. Elles reflètent aussi les conventions et les standards propres à chaque univers socio-professionnel. Les productions issues des données peuvent être des applications, des services à destination d'un large public (développeurs), des systèmes d'information et d'interfaces destinés aux clients (data scientists) ou des informations destinées aux citoyens (data journalists).

Au-delà des caractéristiques professionnelles, chaque réutilisateur cherche à transformer en une information compréhensible des données brutes, hétérogènes, difficilement lisibles par celui qui ne les a pas produites. Cette transformation est le fruit des activités collectives basées sur une chaîne de traitement comprenant la collecte et le stockage, l'exploration, la compréhension et l'analyse des données, la transformation et enfin, l'exploitation/implémentation des données : développement ou modélisation. Nous avons figuré ces étapes dans l'ordre linéaire et chronologique tout en étant conscientes du fait qu'à l'intérieur d'un système de traitement de données, toutes les étapes dépendent des autres et que des aller-retours peuvent s'opérer entre différentes étapes.

2.1 La valeur ajoutée par les éditeurs des données

Les éditeurs des données occupent une place importante dans l'écosystème des données car ils contribuent à leur visibilité et à leur accessibilité et permettent à une donnée d'amener de la valeur dans une chaîne de traitement de données.

Guidés par les thématiques traitées, les réutilisateurs recherchent souvent des données par facettes et mots-clefs dans des portails qui référencent les données au rang national (par exemple, datagouv, portails OD métropolitains) ; avec des moteurs de recherche généralistes (Google, Yahoo) et enfin, dans les bibliothèques de données partagées librement (par exemple, GitHub). Toutefois, retrouver une bonne donnée et savoir qu'elle existe est toujours un véritable défi, surtout si elle relève d'un domaine spécifique ou concerne un sujet sensible, comme par exemple, les détails du budget d'une mairie. La disponibilité d'un jeu des données ne conditionne pas sa réutilisabilité. Produites par des administrations et des collectivités en fonction de leurs compétences et pour leurs besoins spécifiques, les données ouvertes peuvent rarement satisfaire les réutilisateurs issus d'autres univers socioprofessionnels sans une adaptation des jeux de données à ce nouveau contexte.

La valeur de la donnée se définit surtout par rapport à sa qualité, sa fiabilité et son interopérabilité. En effet, la qualité de la donnée est nécessairement évaluée par rapport à sa mise à jour et à sa continuité qui garantissent la pérennité du service informationnel basé sur ces données. A son tour, la fiabilité de la donnée renvoie à la vérification de la donnée et sa qualification à travers une documentation spécifiant ses conditions de production et ses limites. Les données ouvertes comportent souvent des erreurs, des anomalies ou des pannes, ce qui dissuadent certains réutilisateurs : « Les données open data ne sont pas encore très utilisées par les entreprises. Elles n'ont pas encore une qualité suffisante pour être intégrées aux algorithmes prédictifs » (Entretien avec un consultant formateur en data science).

Au-delà des problèmes de qualité et de fiabilité, il se pose également la question d'interopérabilité des données. En effet, les données ouvertes issues de différents producteurs ne sont ni décrites ni indexées de la même manière malgré les injonctions législatives dans la lignée des directives européennes « ISP » et « INSPIRE », qui affirment l'importance des standards ouverts, à savoir des formats interopérables aux spécifications techniques disponibles d'une manière publique et sans restriction d'accès ni de mise en œuvre. Les choix des normes et des standards de publication des jeux de données ouvertes et des métadonnées qui les accompagnent ont un impact décisif sur les

modes de recueil et de traitement des données et sur leur valorisation dans de nouveaux contextes socioprofessionnels. La normalisation suppose à intégrer les données produites par de différents services métropolitains dans un système d'information « qui permet de transformer des données brutes, hétérogènes, difficilement lisibles par celui qui ne les a pas produites en une information visible et compréhensible » (Flichy, 2013). Toutefois, les standards de l'open data sont imbriqués aux standards et normes traditionnellement utilisés pour la production des données dans le cadre de différents domaines de compétences des métropoles. La démarche de l'open data nécessite à vérifier si les formats traditionnellement utilisés sont ouverts et interopérables. A défaut, elle oblige à convertir les documents d'un format « propriétaire » en un format « standard ». C'est le cas, par exemple, des documents PDF dont l'usage est répandu au sein de plusieurs services métropolitains.

De fait, les méthodes de classement, de référencement, d'indexation et les ontologies utilisées par les éditeurs des données conditionnent l'exploitation et la réutilisation des données. Les fournisseurs de plateformes cherchent à proposer des solutions « clé à mains, …qui vulgarisent, qui apportent à leurs utilisateurs la capacité justement à transformer une donnée brute en donnée réutilisable, visualisable » (Entretien avec un fournisseur de portail OD). Cependant, l'étude comparée des portails OD métropolitains que nous avons conduite en 2017 (Paquienséguy, Dymytrova, 2017) relevait une forte hétérogénéité des données ouvertes publiées avec à la fois des formats ouverts comme par exemple JSON ou CSV et des formats propriétaires comme Shapefile, développé par ESRI. Pour pouvoir répondre à divers types de réutilisation et satisfaire à la fois des utilisateurs professionnels et amateurs, certains éditeurs de données multiplient des formats des jeux de données : « Chaque communauté a ses formats favoris. On a toujours un fichier Excel, un fichier ODS pour le commun des mortels, des fichiers KML pour qu'ils puissent le glisser dans Google Maps ou Google Earth, et après des formats plus spécifiques liés à des métiers particuliers » (Entretien avec un chef du projet open data au sein d'une collectivité).

Le fait que les réutilisateurs professionnels souhaitent avoir « une donnée élaborée et maîtrisée par le fournisseur » (Entretien avec un développeur fondateur d'une start-up) rend important le nettoyage et la qualification des données conduits par des éditeurs des données avant leur mise à disposition via des portails dédiés. Les portails OD rendent bien visible travail des éditeurs des données car facilitent réellement la recherche et la récupération des données par des développeurs, des data scientists et des data journalists.

2.2 Développeurs : la valeur ajoutée couplée au désir de solution innovante

Les développeurs produisent des applications web et mobile à destination des clients ou d'un large public. Leur travail se base sur une chaîne de traitement qui comprend les étapes suivantes (Fig.1).

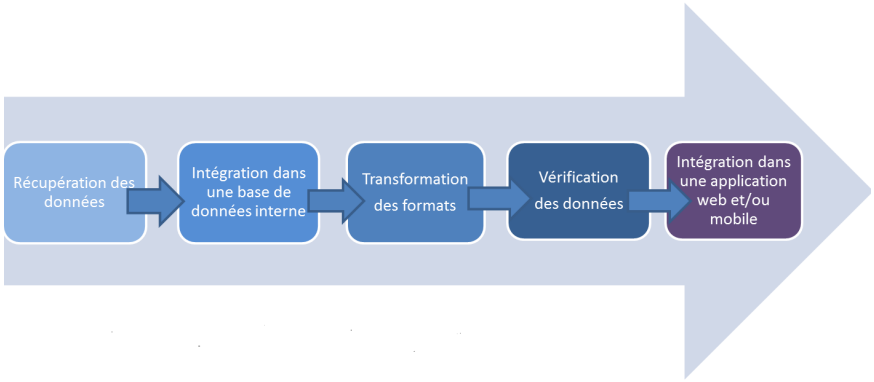

Figure 1. Chaîne de traitement des données par des développeurs

Les développeurs qui évoluent au sein des start-up s'intéressent beaucoup aux données ouvertes pour développer des services utiles aux citoyens-consommateurs. Ils utilisent des données négociées avec les partenaires et des données collaboratives, créées et partagées par les utilisateurs *via* des librairies ou des plateformes collaboratives. Pour eux, l'open data ne constitue pas la source principale d'information. Par ailleurs, plusieurs professionnels interrogés constatent l'absence des données ouvertes spécifiques dont ils auraient besoin pour développer leurs propres produits. Par exemple, les données relatives aux horaires d'ouverture des établissements publics culturels (musées, théâtres, etc.) ou des informations pratiques concernant les conditions d'accès à ces endroits (tarifs, accès handicapé) sont pour l'instant manuellement saisies par les développeurs qui souhaitent les intégrer dans leurs applications.

Les formats initiaux, en particulier quand il s'agit de formats propriétaires et la structuration des jeux de données empêchent souvent la réutilisation. Pour qu'ils puissent interagir facilement avec les données en les recherchant et les récupérant automatiquement, les développeurs ont besoin d'un ensemble de fonctions logicielles (Application Programming Interface - API) appelées depuis l'extérieur de l'application qui les expose. Quand l'API est absente, la récupération et le traitement des données demandent beaucoup de temps et d'efforts d'adaptation. Cependant, toutes les API ne garantissent pas aux développeurs la qualité d'accès aux données. Les problèmes récurrents sont liés aux surcharges provoquées par l'ouverture publique d'une API ou aux dysfonctionnements lors de l'interrogation de certaines données en temps réel, comme c'est le cas à chaque fois qu'il est question d'horaires par exemple.

Plusieurs développeurs interrogés affirment que le format le plus utilisé actuellement est JSON (JavaScript Object Notation) : « Tout le monde utilise JSON aujourd'hui parce que c'est très simple et tous les langages d'information ont quasiment par défaut une librairie qui permet de lire JSON. Du coup, le taux d'adoption est énorme, parce que suffisamment simple pour que tout le monde travaille avec » (Entretien avec un développeur 1).

D'une manière générale, quels que soient les formats de données, les développeurs savent les manipuler et les transformer pour les intégrer à leurs propres bases. La présence de la documentation expliquant l'implémentation du format dans un langage est toutefois primordiale : « On perd un temps fou à rechercher des informations. La documentation c'est 50% du job... » (Entretien avec un développeur 2).

La chaîne de traitement des données des développeurs, facilite la capacité d'innovation (nouveaux services et nouvelles opportunités), mais aussi la capacité de disruption (faire différemment et plus efficacement), parfois à partir de données identiques. Au final, l'appropriation des données qui résulte de leur traitement favorise et précipite leur ré-exploitation.

Pour illustrer cette chaîne de traitement, prenons l'exemple de l'application Géovélo, développée à partir de 2013 par la Compagnie des Mobilités. Cette application d'aide à la navigation permet aux cyclistes de plusieurs villes françaises de calculer un itinéraire en temps réel et de s'informer sur les aménagements cyclables, stations vélo ou encore disponibilités des places. La chaîne de traitement des données commence dans ce cas par l'initialisation des données, à savoir un regroupement de toutes les données, quelle que soit leur nature. Au début de chaque nouveau projet, les développeurs font une comparaison entre ce qui peut exister sur le portail métropolitain de l'OD et ce que les collectivités peuvent avoir chez eux (données fermées). Cela amène la start-up à travailler directement avec les SIG de certaines collectivités pour avoir des données non disponibles via les portails OD. En réutilisant les données des collectivités, par exemple les données liées aux aménagements cyclables, l'équipe Géovélo contribue à améliorer leur qualité grâce à un vrai travail de terrain ce qui permet de compléter les données absentes ou corriger certaines informations. Une piste cyclable peut par exemple s'avérer une bande cyclable ce qui aura un impact sur le déplacement d'un cycliste. Une fois les données initialisées, elles sont retraitées et filtrées en fonction des objectifs visés, à savoir calcul des itinéraires en temps réel et cartographie. Les données sont ensuite mises en cohérence pour pouvoir bien fonctionner ensemble une fois elles seront reliées. Géovélo s'appuie sur les données géographiques issues d'OpenStreetMap, le wiki cartographique mondial qui fournit les données géographiques réutilisables sous licence libre ODbL et alimente lui-même ce projet avec ses propres données, issues entre autres des mises à jour signalées par des usagers-cyclistes. Grâce à l'exploitation et au croisement de différents jeux de données, l'application s'inscrit dans un cadre d'innovation sociale répondant aux besoins des déplacements doux et de la mobilité durable.

2.3 Les besoins des clients, un élément central de la création de valeur par des data scientists

Les data scientists mobilisent des modèles statistiques et mathématiques pour produire de l'information avec une visée d'aide à la décision afin de répondre précisément aux demandes de leurs clients, qui sont principalement des grandes entreprises. Pour cela, ils exploitent différents modèles (par exemple celui d'optimisation, de simulation ou de prédiction) et différentes méthodes (par exemple, catégorisation automatique, analyse comportementale, ciblage ou machine learning). Les données ouvertes constituent pour eux une source mineure, convoquée pour enrichir celles fournies par des clients ou pour croiser plusieurs sources de données entre elles.

La chaîne de traitement des données se présente pour les data scientists de la façon suivante (Fig. 2).

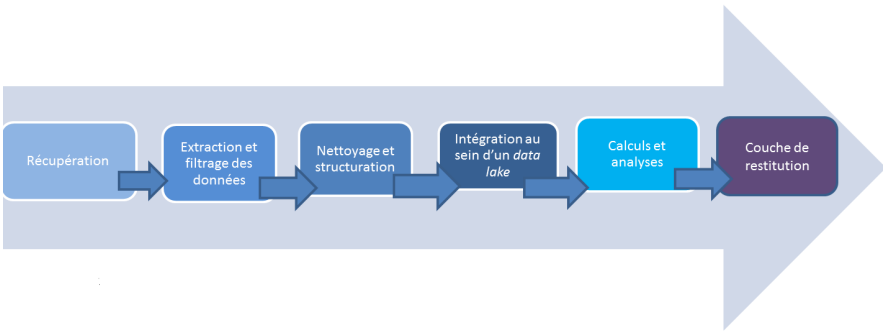

Figure 2. Chaîne de traitement des données par les data scientists

Le filtrage de valeurs et la statistique occupent une place importante dans la chaîne de traitement des données par des data scientists. La transformation des données et leur mise en forme (filtrage en fonction des éléments recherchés, élimination de redondances, structuration et transformation dans les formats utilisés) sont des étapes particulièrement chronophages.

En raison de très gros volumes de données traités, les data lake (référentiels de stockage) jouent un rôle primordial dans leur travail : « Ce sont des plateformes qui permettent de traiter des données massives avec une rapidité importante et de traiter des données de tout type : structurée, semi structurée ou non structurée et on va formater de manière à les rendre propres à l'analyse pour ensuite offrir une couche de restitution aux différents métiers dans l'entreprise qui vont être amenés à exploiter la data » (Entretien avec un data scientist 1).

Pour le stockage des données, ils mobilisent les technologies Apache, comme le socle d'application open source Hadoop pour construire et modéliser les différents entrepôts de données et le modèle de programmation MapReduce qui permet d'accéder aux big data ainsi stockées. Celui-ci permet un traitement distribué des big data : il s'agit de découper les traitements complexes en ensembles de traitements pouvant être réalisés sur des machines séparées, de les piloter à distance et de ré-agréger les résultats afin de limiter les problèmes de scalabilité. En effet, les outils particulièrement appréciés sont ceux qui répondent aux exigences de scalabilité en permettant de traiter des volumes de données beaucoup plus rapidement sans remettre en question les performances du système.

Pour illustrer le fonctionnement de cette chaîne de traitement, nous mobiliserons les propos d'un data scientist et fondateur d'une entreprise de valorisation des données et spécialiste du traitement haute performance de grands volumes de données et d'algorithmes auto-apprenants (machine learning). L'interviewé précise d'abord que la recherche des données se fait en fonction de la problématique des clients et prend de moins en moins de temps grâce aux différents portails qui référencent les données, comme par exemple datagouv, Google, Yahoo ou GitHub. Toutefois, les données ouvertes sont utilisées « au coup par coup » sans mise en production. Car elles n'ont pas encore une qualité suffisante pour être intégrées aux algorithmes prédictifs des entreprises. Le data scientist interviewé travaille beaucoup sur des données issues des systèmes d'informations de ses clients, comme par exemple des données liées à leurs activités, métiers ou des données d'activité industrielles, d'activités marketing et d'activités gestion des ressources humaines. C'est surtout la transformation des données et leur mise en forme qui constituent des tâches les plus chronophages. S'il existe des outils qui facilitent la transformation et la mise en forme des données, ils arrivent vite à une certaine saturation et rendent nécessaire le recours au codage pour pouvoir récupérer la donnée et la retransformer. Une fois la donnée récupérée et transformée, elle est intégrée au

sein d'un data lake où l'on peut lancer les calculs en faisant les liens entre les données qui viennent des sources différentes et qui sont souvent de type différent (structurée, semi structurée ou non structurée). Enfin, les résultats des calculs sont intégrés dans des logiciels de visualisation des données pour proposer une couche de restitution compréhensible par des clients.

Le volume et la diversité des données traitées, mais aussi des chaînes de traitement basées sur des approches distribuées des calculs permettent aux data scientists de mieux répondre aux nouvelles attentes des clients.

2.4 La médiation des données, un axe structurant de la création de valeur selon les data journalists

Les data journalists utilisent les données pour mener des enquêtes, vérifier des informations ou mettre en avant certaines données en apportant une plus-value éditoriale. Les productions finales prennent alors la forme de mises en scène des données à travers des visualisations, des cartographies, des graphiques et des applications interactives. La priorité des data journalists est de se faire comprendre d'un large lectorat, ce qui les oblige à soigner les interfaces utilisateurs.

Les données ouvertes constituent une source des données parmi d'autres pour les data journalists, qui sont souvent à la recherche des données inédites qui n'existent pas ou qui ne sont pas publiques. Comme ils souhaitent disposer de données brutes pour être au plus près de la source, ils recourent souvent à des données collectées d'une manière journalistique auprès d'un réseau de contacts et à des données crawlées ou scrapées sur le web par des robots d'indexation.

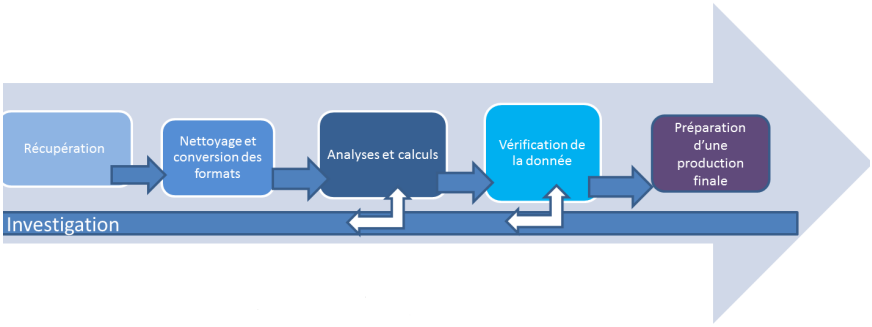

Figure 3. Chaîne de traitement des données par les data journalists

Dans cette chaîne (Fig. 3), la compréhension et le nettoyage des données sont des étapes déterminantes : « Il y a quelque chose de très ingrat là-dedans, par exemple, à passer des heures pour nettoyer un tableur, des heures pour coder quelque chose, de tester, d'essayer de débuguer, des étapes préliminaires avant la visualisation des données, côté méta qu'on ne voit pas quand voit la production achevée » (Entretien avec un data journalist1).

Le traitement des données s'accompagne toujours d'un travail d'investigation traditionnel mené en parallèle car « Il ne faut pas penser que la vérité est à l'intérieur du tableau, il y a toujours un moment d'enquête traditionnelle à faire pour corroborer ce que racontent les données » (Entretien avec un data journalist 2).

Pour illustrer la chaîne de traitement propre à des data journalists, prenons des visualisations de données comme par exemple des cartographies électorales réalisées par l'un de nos interviewés. Dans ce cas, les données historiques proviennent de la plateforme nationale datagouv.fr alors que les données en temps réel concernant des

résultats du vote sont fournies par le Ministère de l'intérieur et mises à jour toutes les 15 minutes. Les data journalistes aspirent ou « scrapent » ces données et transforment le format XML en tableur Excel avec le logiciel OpenRefine. Le fichier Excel est réinjecté ensuite dans une feuille de calcul Google et celle-ci est par la suite transformée en JSON. Là encore, la préparation des données représente une étape chronophage, surtout parce qu'elle comporte également le besoin de vérification des données et la résolution des problèmes liés par exemple à l'encodage des caractères. Avant de passer aux visualisations cartographiques, il est nécessaire de mobiliser un carnet d'adresse d'experts pour pouvoir vérifier et mettre en relief certaines informations. Par exemple, après avoir interviewé un chercheur, spécialiste du FN, le data journaliste en question a réalisé une autre visualisation qui reflétait la différence entre sympathisants et militants du FN. Il s'agit d'une cartographie de toutes les communes de l'Alsace, avec en vert les communes où le score du FN était inférieur au score national (30 communes), une autre couleur pour les communes où le score était plus important que le score national (beaucoup de violet) et enfin, une couleur pour les communes où le FN présentait une liste (6 ou 7 communes). Cette visualisation montrait qu'il y avait des communes qui votaient pour le FN d'une manière plus soutenue ou plus massive qu'au niveau national mais qui n'avaient pas de militants prêts à figurer sur une liste municipale.

Le travail des data journalists montre que la valorisation des données consiste aussi bien dans la maîtrise des outils et des technologies que dans la recherche des angles pertinents de croisement et d'analyse permettant de passer des données aux informations.

3. L'outillage, élément incontournable de la création de valeur

Au-delà des compétences professionnelles clés de la réussite des chaînes de traitement des données, notre enquête a également relevé l'importance de l'outillage. En effet, la démultiplication des données ouvertes et des big data accroît le besoin d'outils et de techniques efficaces pour en tirer le profit. Pour chaque catégorie d'acteurs professionnels interviewés, nous avons relevé des formats, des langages, des logiciels et des API les plus cités (voir la figure 4). Sans évoquer la prédominance des outils d'origine américaine, les solutions citées se différencient entre elles en termes de modes d'accès et d'exploitation. En effet, les logiciels open source y côtoient des solutions propriétaires. Certaines solutions nécessitent une véritable maîtrise du code, d'autres sont plutôt clé en main pour permettre aux professionnels moins expérimentés de créer des visualisations des données ou des applications mobiles plus facilement et rapidement.

Figure 4. Solutions professionnelles citées par les interviewés.

Développeurs	Data scientists/analysts	Data journalists
Formats et langages : JSON, CSV, GeoJson, Shape, KML, Python, GeoPy, SQL, Java Script, PHP, PHP Symfony 3	Formats et langages : XML, Scala, Java, Apache Parquet, langage et bibliothèques Python	Formats et langages : Excel, XLS, JSON, CSV, KLS, JavaScript, Python
Logiciels et API : OpenStreetMap, Google Map, Postman de Google Chrome, FME, GeoServer, GeoNetwork, Csvkit, Addok	Logiciels et API : R, QJIS, Hadoop, Power BI (Buisiness Intelligence) de Microsoft, Cplex d'IBM, Talend, MongoDB, Elastic Search,	Logiciels et API : OpenRefine, Google Spread Sheets, Google Fusion Tables, Infogram, ChartBlock, Gephi, Datawrapper, CARTO, QGIS, Mapbox, D3

	Dataiku, Scope, Spark, Qlik, Tableau	

Si la panoplie des outils professionnels est très large et varie selon les métiers, nous constatons une transversalité de l'usage des tableurs, du logiciel R, du langage de programmation Python et du SQL (Structured Query Language).

Les développeurs exploitent systématiquement des données géographiques avec des services de géocodage comme Addock, Géoserver ou GeoNetwork ou encore l'API de Google Map car les services qu'ils développent ont un fort ancrage territorial. La palette des outils professionnels des développeurs interrogés dépend des fonctionnalités et des thématiques qu'ils proposent dans leurs applications.

Les data scientists sont bien outillés à toutes les étapes de la chaîne de traitement des données. Pour l'extraction de la donnée, ils utilisent les logiciels de la famille ETL (Extract Transform Load) qui effectuent des synchronisations massives d'information entre sources de données. Parmi ces logiciels, certains citent une solution française open source Talend, qui couvre un grand nombre des besoins en intégration et transformation des données. Un autre outil incontournable mentionné est le logiciel de traitement statistique R avec sa palette étendue de fonctionnalités graphiques. Ce qui distingue les data scientists d'autres professionnels étudiés est leur compétence en traitement distribué de big data. Les outils particulièrement appréciés par les data scientists sont ceux qui permettent les calculs distribués comme les socles d'application Hadoop et Spark.

Les data journalists mobilisent beaucoup d'outils, très variés comme les logiciels d'édition et de présentation de tableaux. Certains travaillent avec Excel, d'autres lui préfèrent Google Spread Sheets pour partager les feuilles de calcul. Pour le nettoyage et la mise en forme des données, ils emploient OpenRefine afin de remédier au problème récurrent lié à l'encodage (présence/absence des caractères spéciaux, par exemple). Ce logiciel satisfait bien le besoin des journalistes d'avoir des données brutes n'ayant pas subi de traitements réduisant leur potentiel d'information : « Je préfère des fichiers où il n'y a pas de trop de simplification, avec des erreurs, des trous, des remarques qui me permettent d'avoir plus de précision, plutôt que les fichiers trop propres, nettoyés où on a supprimé le niveau communal et on a tout mis au niveau cantonal » (Entretien avec un data journalist 3).

D'une manière générale, les data journalists apprécient beaucoup les bibliothèques logicielles car elles fournissent des échantillons de codes et des exemples d'utilisation qui leur servent de base pour le développement des applications. Certains citent le langage open source Python qui doit son succès à une grande quantité de bibliothèques, créées et maintenues par une large communauté d'usagers.

L'évolution constante des technologies risque de rendre rapidement archaïques, voire obsolètes certaines solutions professionnelles que nous évoquons. Toutefois, nos analyses montrent l'importance de penser les chaînes de valeur de données en termes d'« ingénierie système » avec une attention particulière accordée à l'écosystème logiciel et à une dynamique communautaire.

En effet, les données sont valorisées grâce à l'articulation des compétences « hardware », assurant la gestion des réseaux de ressources distribuées, des compétences « systèmes informatiques », assurant un bon fonctionnement du système d'exploitation, des compétences logicielles applicatives liées à l'exploitation des solutions techniques et enfin, des compétences algorithmiques. La diversité des outils et des logiciels mobilisés dans les chaînes de traitement de données montre qu'elles sont loin d'être industrialisées et fonctionnent à partir d'ajustements, de détournements, de bricolages et dans une logique *ad hoc* spécifique au courtage informationnel. De leur efficacité et de leur efficience dépendent la rapidité de la réalisation et la qualité des services basés sur les données.

En même temps, la création de valeur à partir des données est fortement marquée par une dynamique communautaire. Les librairies et référentiels ouverts (par ex., GitHub, OpenStreetMap ou Dbpedia) facilitent le travail de nombreux réutilisateurs professionnels qui tiennent à leur tour à partager les données obtenues par leurs calculs et les analyses ou les codes créés avec d'autres, dans l'esprit du mouvement open source. En effet, la préférence envers des logiciels open source est nette dans notre panel. D'abord, par la proximité idéologique entre la communauté des réutilisateurs de l'open data et le mouvement open source, ensuite, par le caractère émergent des solutions développées, avec une distribution du travail de test et de validation qui leur permet une réponse systémique efficace à des demandes variées. Dans tous les cas, l'opposition entre les solutions libres et les solutions propriétaires structure la compréhension de la valeur ajoutée créée, dans les propos des professionnels interrogés.

4. Discussion

Notre enquête de terrain souligne deux points qui méritent discussion sur la base des pratiques professionnelles des personnes interrogées et des enjeux qui s'en dégagent.

Tout d'abord, celui de la place des données ouvertes dans l'écosystème général des données travaillées ou produites par ces professionnels, autrement dit la place des open data dans les big data dont la valeur et la variété sont ici les caractéristiques les plus centrales. En effet, les données ouvertes participent des big data dont elles tentent de vérifier certaines ces deux caractéristiques parmi les célèbres 5V : Volume, Vitesse, Variété, Valeur et Véracité. La question centrale ici est d'abord celle de la valeur, car c'est l'élément premier de distinction entre les données ouvertes et les big data. D'ailleurs, les données ouvertes métropolitaines tendent à changer de nom dans plusieurs écosystèmes métropolitains pour devenir données partagées ou, mieux, données publiques ; lesquelles s'opposent bien entendu à données privées. Les premières, produites par les administrations françaises seraient d'intérêt général et soumises à une licence d'un nouveau genre qui s'y rapporte (LRDIG)[2] ; les deuxièmes, issues d'acteurs du secteur privé seraient source de valeur et soumises à exploitation commerciale. Comme nous l'avons montré, « la fragmentation territoriale des données ouvertes et les spécificités locales des portails métropolitains freinent, […] leur exploitation à l'échelle industrielle, d'où la présence d'autres types d'acteurs, plus souples, moins systématiques qui utilisent des données de niche (quelques jeux seulement) pour un développement ciblé et limité, à un territoire dans la plupart des cas. » (Paquienséguy, 2020, p.11). La valorisation des données ouvertes, ou publiques se pose principalement aujourd'hui en termes de services, de data-services, c'est pourquoi les réutilisateurs sont un élément-clé des données métropolitaines. Ces professionnels des données développent soit des services à destination de certaines catégories d'acteurs. Ils assurent ainsi une fonction, sélective, de médiation des données (Dymytrova, 2020) et transforment le chiffre en information utile *via* une typologie de licences qui posent déjà le contexte économique et juridique du réemploi des données publiques ouvertes. Cet espace économique de la métropole correspond à une nouvelle phase de son développement économique à considérer comme un élément stratégique du marketing territorial et de son évolution vers la smart city. Entre fluidité et surveillance[3], les données de mobilité urbaine sont emblématiques, tant par les enjeux environnementaux qu'elles portent que par les questions de la place de l'usager et de la protection de sa vie privée qu'elles posent. C'est ici un point de croisement important de voir la dimension éthique s'imposer aux données ouvertes devenues massives et stratégiques pour un développement économique durable du territoire. S'y retrouvent d'une part la protection

[2] https ://www.grandlyon.com/delibs/pdf/Conseil/2019/09/30/DELIBERATION/2019-3724.pdf, consulté le 7 juin 2020

[3] Le Forum InOut de Rennes était consacré à cette question en mars 2019 : « Mobilité urbaine, le partage de données entre fluidité et surveillance ».

des données personnelles de mobilité ou de consommation de services publics (bibliothèques ou piscines par exemple) et d'autre part l'éthique de la réutilisation des données entre ruée vers l'or et intérêt général qui ouvre sur une dimension critique de la situation constatée dans notre enquête plus : value *vs* partage ; données privées *vs* open data ; big data *vs* privacy.

Ensuite, celui de la place ou de la source de l'innovation dans les pratiques de production et de réutilisation des données et c'est ici vers les producteurs de données ouvertes ou partagées qu'il faut se tourner, à partir des métropoles qui sont le maître d'œuvre. Leader puissant, sur les données de mobilité urbaine, mais pas seulement, certaines, dont Lyon, Montréal ou Boston sont particulièrement actives à travers le programme Movin'On[4], et œuvrent à la mise en place de bonnes pratiques autour des données métropolitaines de plus en plus hétérogènes en termes d'acteurs (privés, publics, industriels et start-up, makers de tous ordres). C'est donc bien dans la dimension organisationnelle et structurelle que se niche en partie l'innovation portée par les données ouvertes d'un territoire comme le souligne également l'étude ethnographique d'Antoine Courmont (2019). Cependant, elle se fait à marche forcée dans la plupart des cas car la position centrale de la métropole impose à ses interlocuteurs publics comme aux entreprises du territoire concernées d'entrer dans le modèle de production et de réutilisation des données qu'elle propose. La dernière campagne du GrandLyon à ce sujet est très claire : « all data makers »[5]. L'innovation ne vient pas donc des outils ou des formats utilisés, ni même des pratiques professionnelles qui, comme on l'a vu, ne peuvent se développer uniquement autour ou à partir des données ouvertes, toujours associées à d'autres types de données avant de se transformer en data-services ou en information. Elle se présente vraiment comme structurelle et porte les germes d'un « modèle urbain ». Ainsi, les portails open data et leurs utilisateurs participent-ils à un processus d'innovation urbaine (Lamarche, 2003) riche de plusieurs options : la ville durable, la ville intelligente, la ville créative, la ville participative, etc. Les enjeux de l'exploitation des données ouvertes dépassent donc largement leur propre cadre (législatif, technique et économique) pour s'inscrire dans une démarche de projet politique de développement du territoire bien plus conséquente, démarche dans laquelle la place faite ou laissée au citoyen peut considérablement varier, nous l'avons vu. Dans une moindre mesure, des modèles de métropoles data-driven se profilent comme c'est déjà le cas à Hangzhou, Shenzhen ou Toronto - dont le projet Skylabs a cependant été interrompu. Mais d'autres innovations urbaines sont également à l'œuvre, sur de vraies logiques open data comme, dans des métropoles qui développent leur portail open data en open source (Lyon, Montpellier, Nantes, Montréal, Séoul, Boston etc.) et interagissent avec l'écosystème *via* des laboratoires d'innovation urbaine qui favorisent la collaboration. Les grandes questions soulevées sont alors celles des bonnes pratiques dans cet écosystème de plus complexe, avec, dans certains cas l'objectif de faire de la donnée métropolitaine, un bien commun substrat de services partagés.

Commenté [I1]: A intégrer dans la biblio

Conclusion

Le modèle de chaîne de valeur permet d'analyser les réutilisations des données comme une série d'opérations contribuant à produire de nouveaux points de vue et des informations utiles, tout en générant de la valeur. La valeur n'est pas intrinsèque aux données ouvertes, mais provient des explorations et des transformations de ces données par divers acteurs de l'écosystème. En effet, les chaînes de traitement identifiées montrent comment le sens vient aux données ou plus précisément comment les données sont progressivement analysées et traitées pour permettre

[4] https://movinon-lab.michelin.com/lab/s/movinon-lab?language=fr
[5] https://download.data.grandlyon.com/files/grandlyon/Plateforme_Data_GrandLyon_MetropoledeLyon_EN.pdf, consulté le 7 juin 2020

l'extraction d'information (vérification, agrégation, validation), puis, dans certains cas, devenir connaissances adaptées à un nouveau cadre social (analyses, calculs) et enfin pour être transformées en services, qui font apparaître la valeur ajoutée car peuvent faire l'objet de monétisation (couches de restitution, applications). Or, la valeur des données ouvertes ne peut plus être uniquement pensée en termes économiques. Les réutilisations de ces données par des data journalists offrent un exemple de valeur sociale des données qui contribuent à l'information (Dymytrova, 2018) et à l'empowerment des citoyens (Goëta, Mabi, 2014 ; Badouard, 2017).

Si les chaînes de traitement analysées recourent à des données et des outils particuliers, elles illustrent les liens entre les acteurs publics et privés de l'écosystème métropolitains. La création de la valeur ajoutée dépend principalement de la qualité des données ouvertes (mise à jour régulière, présence de métadonnées et de la documentation, interopérabilité de formats). La vérification et la qualification des données sont ainsi des éléments importants dans la chaîne de traitement pour l'ensemble des professionnels étudiés, d'où l'importance de la confiance accordée aux données et la nécessité pour les producteurs d'expliciter leurs méthodologies. Nous retrouvons là de façon nette deux éléments caractéristiques du courtier ou de l'entreprise de courtage : 1/ la valorisation/monétarisation du travail de l'intermédiaire, dans le cas du data journalisme mode de rémunération reste indirect, contrairement aux deux autres catégories professionnelles ; 2/ le recours à des outils et routines de développement ; la proximité avec le courtage informationnel renforce encore la nécessité de porter la méthodologie à connaissance de l'usager ou du destinataire final du service ou de l'information.

Références

Badouard R. 2017. Open government, open data : l'empowerment citoyen en question, in : Ouvrir, partager, réutiliser. Regards critiques sur les données numériques, Paris, Éd. de la Maison des sciences de l'homme.

Bardou-Boisnier S., Pailliart I. 2012. Information publique : stratégies de production, dispositifs de diffusion et usages sociaux. Les Enjeux de l'information et de la communication, 13 (2), p. 3.

Bouquillon P., Miège B., Mœglin P. 2013. L'industrialisation des biens symboliques. Les industries créatives en regard des industries culturelles. Presses universitaires de Grenoble, p. 82.

Boustany J. 2013. Accès et réutilisation des données publiques : État des lieux en France, Les Cahiers du numérique, 9 (1), 21-37.

Courmont A. 2019. Ce que l'open data fait à l'administration municipale. La fabrique de la politique métropolitaine de la donnée. Réseaux, 6 (218), p. 77-103.

Courmont A. 2015. La plateforme de diffusion de données, un modèle de gouvernement urbain ?, dans : Évelyne Broudoux éd., Big Data - Open Data : Quelles valeurs ? Quels enjeux ? Actes du colloque « Document numérique et société », Rabat, 2015. Louvain-la-Neuve, De Boeck Supérieur, « Information et stratégie », p. 85-95.

Curry E. 2016. The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches, in: Cavanillas J., Curry E., Wahlster W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham.

Dymytrova V., Larroche V., Paquienséguy F. 2018. Cadres d'usage des données par des développeurs, des data scientists et des data journalistes. Livrable n°3. EA 4147 Elico. Accès : https://hal.archives-ouvertes.fr/hal-01730820/document.

Dymytrova V., Paquienséguy F. 2017. La réutilisation et les réutilisateurs des données ouvertes en France : une approche centrée sur les usagers, Revue Internationale des Gouvernements Ouverts, 5, 117-132.

Dymytrova V. 2018. Data journalisme, entre pratique créative innovante et nouvelle médiation experte ? Une analyse conjointe des discours et des productions journalistiques, XXIe Congrès de la SFSIC « Création, créativité et médiations », 13-15 juin 2018, MSH Paris Nord. vol.3. Accès : www.sfsic.org/attachments/article/3280/Actes%20vol%203%20-%20congr%C3%A8s%20SFSIC%202018.pdf.

Flichy P. 2013. Rendre visible l'information. Une analyse sociotechnique du traitement des données, *Réseaux*, 2 (178-179), 55-89. Accès : https://www.cairn.info/revue-reseaux-2013-2-page-55.htm.

Goëta S., Mabi C. 2014. L'open data peut-il (encore) servir les citoyens ?, Mouvements, 3 (79), 81-91.

Kitchin R. 2014. The data revolution: Big data, open data, data infrastructures and their consequences, London, Sage.

Labelle S., Le Corf J-B. 2012. Modalités de diffusion et processus documentaires, conditions du « détachement » des informations publiques. Analyse des discours législatifs et des portails open data territoriaux, Les Enjeux de l'Information et de la Communication, 13 (2), p. 210.

Lamarche T. 2003. « Le territoire entre politique de développement et attractivité », Études de communication, 1 (26), p. 9-19.

Larroche V., Vila-Raimondi M. 2015. Urban Data et stratégies dans le secteur des services : le cas de la métropole lyonnaise, in : Big Data - Open Data : Quelles valeurs ? Quels enjeux ?, Louvain-la-Neuve, De Boeck supérieur, 183-195.

Lévesque B. 2007. Une gouvernance partagée et un partenariat institutionnalisé pour la prise en charge des services d'intérêt général. CRISES.

Mœglin P. 2005. Outils et médias éducatifs. Une approche communicationnelle, Presses universitaires de Grenoble.

Noyer J.-M., Carmes M. 2013. Le mouvement « Open data » et les intelligences collectives », in : Les débats du numérique, Paris, Presses des Mines, 137-168.

Paquienséguy F. 2016. Les portails open data au prisme du courtage informationnel : qu'est ce qui se joue pour les métropoles ?, in : Open Data, accès, collectivités territoriales et citoyenneté : des problématiques communicationnelles, Paris, Éditions des Archives contemporaines.

Paquienséguy F. 2020. « L'open data métropolitain en voie de structuration ? » in : Données urbaines et smart cities, Paris, Éditions des archives contemporaines, 2-31

Paquienséguy F., Dymytrova V. 2017. Livrable 1.2 Analyse de portails métropolitains de données ouvertes à l'échelle internationale. [Rapport de recherche]. Équipe d'accueil lyonnaise en Sciences de l'information et de la communication. ⟨hal-01449348⟩.

Turky S., Foulonneau M. 2015. Valorisation des données ouvertes : acteurs, enjeux et modèles d'affaires, in : Big data - Open data : Quelles valeurs ? Quels enjeux ?, Louvain-la-Neuve, De Boeck Supérieur, 113-125.