Des données aux services dans les Métropoles : enquête auprès des réutilisateurs professionnels de données ouvertes

Valentyna Dymytrova, Françoise Paquienséguy

Des données aux services dans les Métropoles : enquête auprès des réutilisateurs professionnels de données ouvertes

Valentyna Dymytrova[1], Françoise Paquienséguy[1]

[1]EA 4147 Elico, Sciences Po Lyon, France

E-mail : valentyna.dymytrova@sciencespo-lyon.fr

Résumé

A partir d'une enquête de terrain menée en France en 2017, cette communication identifie différentes formes de réutilisations 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 journalistes), 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. Cette communication est partiellement issue des travaux de l'ANR OpenSensingCity.

Mots-clés : open data, chaîne de valeur, développeurs, data scientists, data journalistes

Abstract

Based on a field survey conducted in France in 2017, this communication identifies different forms of open data reuse and analyses the processing data chains on which they are relied. By analyzing 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 communication results partially of the research ANR OpenSensingCity.

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

---

1. Introduction

Le processus d'ouverture des données (OD) 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 [1]. Comme le soulignent Noyer et Carmès [2], 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. 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 connaissance et de services dans de nouvelles formes et conditions, configurées par la réutilisation des données ouvertes.

Si les cadres législatif et administratif de l'OD ont été récemment redéfinis par plusieurs lois [3] et par le Plan d'action national 2015-2017, rendre les données ouvertes ne suffit pas à générer leur réutilisation [4]. Celle-ci constitue un processus complexe impliquant plusieurs communautés professionnelles aux divers objectifs stratégiques et économiques et agissant au sein de divers cadres éthiques [5]. 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 [6].

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 » » [7]. 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 l'OD 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 » [8]. Acteurs de cette sous-filière, porteurs du modèle du courtage informationnel [9], 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 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 [10]. 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, au-delàs de 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 des modalités de création de la valeur, 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 [...] ? » [11].

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 ou applications ? 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 journalistes) 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 de 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/analystes, 6 data journalistes, 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 [12].

2. Chaîne de traitement des données et création de valeur

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 journalistes).

Au-delà des caractéristiques professionnelles, chaque réutilisateur est confronté à 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.

2.1 La valeur ajoutée par les éditeurs des données

Guidés par les thématiques traitées, les réutilisateurs interrogés 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. Le problème d'interopérabilité se

pose également puisque les données ouvertes issues de différents producteurs ne sont ni décrites ni indexées de la même manière.

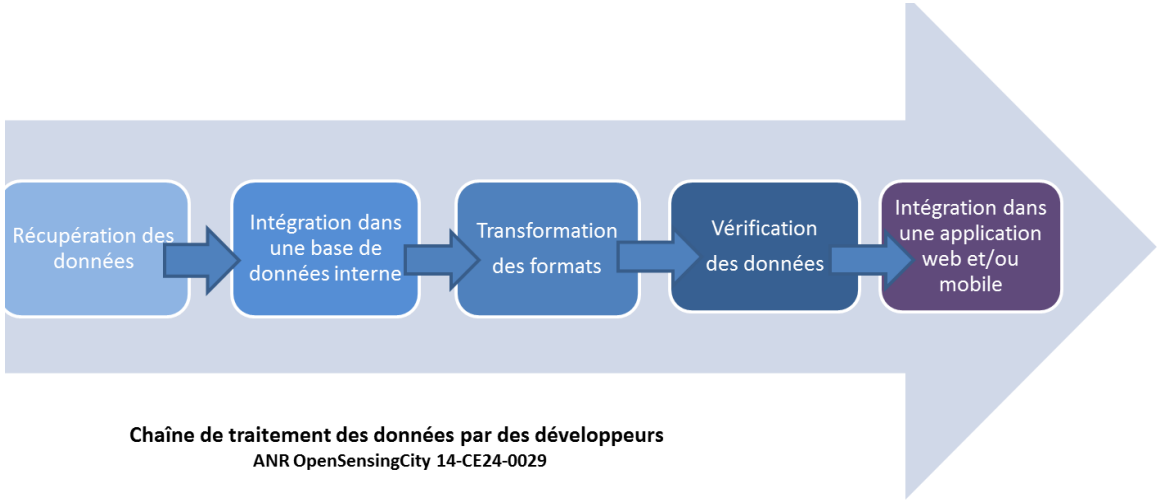
Les éditeurs des données occupent ainsi une place importante dans l'écosystème des données car ils contribuent à leur visibilité et à leur accessibilité. Les méthodes de classement, de référencement, d'indexation et les ontologies qu'ils utilisent conditionnent l'exploitation et la réutilisation des 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).

La suite de l'article présente donc les spécificités des chaînes de traitement de données et des outils professionnels des développeurs, des data scientists et des data journalistes.

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 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 (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 » (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... » (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

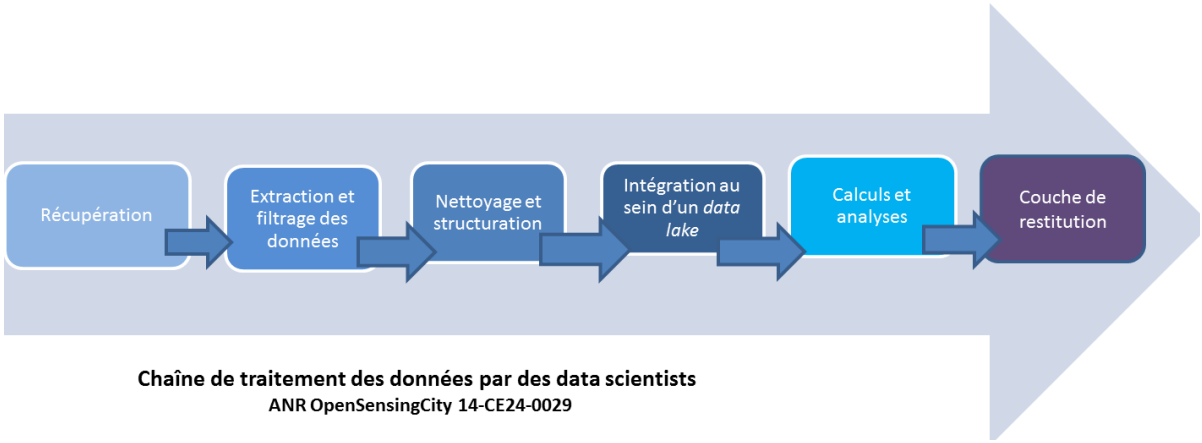
partir de données identiques. Au final, l'appropriation des données qui résulte de leur traitement favorise et précipite leur ré-exploitation.

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 (d'optimisation, de simulation, de prédiction) et différentes méthodes (catégorisation automatique, analyse comportementale, ciblage, machine learning). L'OD constitue pour eux une source mineure de données, convoquée pour enrichir celles fournies par des clients ou pour croiser plusieurs types 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 (référentiel 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 » (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. 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 d'éviter 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.

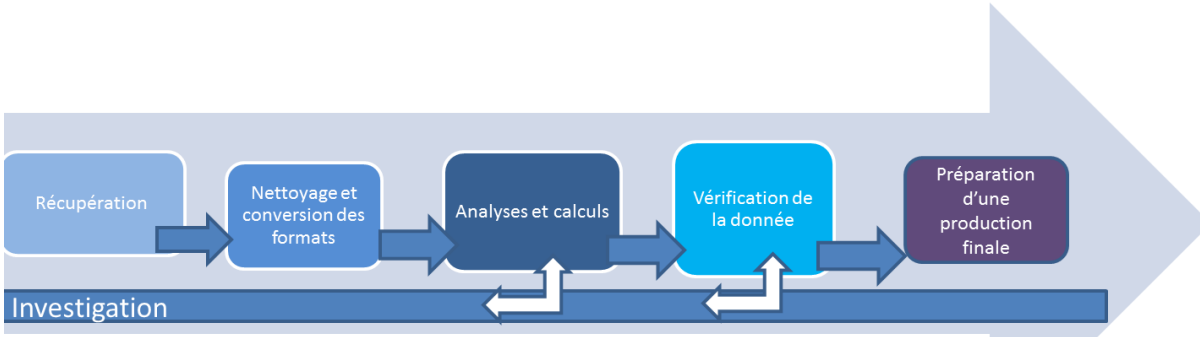
Le volume et la diversité des données traitées, mais aussi des chaînes de traitement basées sur des approches distribuées des calculs permettent aux data scientists de mieux répondre aux nouvelles attentes des clients.

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

Les data journalistes 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 d'articles de presse traditionnels et de mises en scène des données à travers des visualisations, des cartographies, des

graphiques et des applications interactives. La priorité des data journalistes est de se faire comprendre d'un large lectorat, ce qui les oblige à soigner les interfaces utilisateurs.

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



Chaîne de traitement des données par des data journalistes
ANR OpenSensingCity 14-CE24-0029

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 » (data journaliste 1).

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 » (data journaliste 2).

Le travail des data journalistes 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

La démultiplication des données ouvertes et des big data accroît le besoin d'outils et de techniques efficaces pour en tirer le profit. 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 mentionnés est le logiciel de traitement statistique R avec sa palette étendue de fonctionnalités graphiques.

Les data journalistes 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 » (data journaliste 3).

D'une manière générale, les data journalistes 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.

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'OD 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.

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 [13].

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 [14]. En effet, les chaînes de traitement identifiées montrent comment le sens vient aux données ou plus précisément comment les données se transforment progressivement en information (vérification, agrégation, validation), pour ensuite devenir connaissances adaptées à un nouveau cadre social (analyses, calculs) et enfin pour se transformer en services, qui font apparaître la valeur ajoutée car peuvent faire l'objet de monétisation (couches de restitution, applications). Or, la valeur de l'OD ne peut pas être uniquement pensée en termes économiques. Les réutilisations de ces données par des data journalistes offrent un exemple de valeur sociale des données qui contribuent à l'information [15] et à l'empowerment des citoyens [16].

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 de l'OD urbain [17]. 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 journaliste le 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

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

[2] 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.

[3] Notamment, les lois Macron, NOTRe, Valter et Lemaire.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] Dymytrova V. 2018. Les médiations de l'open data au prisme des applications liées à la mobilité, Les Enjeux de l'Information et de la Communication, 19 (2), 81 - 92.

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

[16] 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.

[17] 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.