Cadres d'usage des données par des développeurs, des data scientists et des data journalistes Livrable n°3

Valentyna Dymytrova, Valérie Larroche, Françoise Paquienséguy

elico

Équipe de recherche de Lyon en sciences de l'information et de la communication

Cadres d'usage des données par des développeurs, des data scientists et des data journalistes

Livrable n°3

Rédaction : Valentyna Dymytrova (Postdoc)

Contribution : Valérie Larroche (MCF), Françoise Paquienséguy (PU)

22 janvier 2018

INTRODUCTION

Porté par Elico (EA 4147), ce livrable n°3 est associé à la tâche D2.3 : **«** Results of inquiry of end users' needs and practices, and usage scenarios ». Il présente les résultats de l'enquête auprès des réutilisateurs professionnels des données ouvertes, menée de février à avril 2017. L'objectif de ce travail a été triple :

- mieux comprendre et qualifier le cadre d'usage des données ouvertes par différents professionnels ;
- identifier et analyser la chaîne du traitement des données ;
- faire apparaître leurs besoins en matière d'outils et de technologies nécessaires à la réutilisation des données.

Le livrable qui résulte de cette enquête est organisé en trois parties. La première décrit le contexte et présente l'état de l'art sur les réutilisations des données ouvertes. La seconde explicite la méthodologie déployée dans la conduite de l'enquête et dans l'analyse des résultats. La troisième partie rend compte des résultats de l'enquête en présentant d'abord les spécificités de données utilisées et en analysant ensuite la chaîne de traitement de données par des développeurs, des data scientists et des data journalistes interviewés. Enfin, la synthèse, disponible à la page 41, met en exergue des points à prendre en compte pour favoriser la réutilisation des données ouvertes par des professionnels des données.

1. CONTEXTE ET ÉTAT DE L'ART

En France, le gouvernement et les collectivités territoriales se sont engagés dans la mise à disposition d'informations mais également de données ouvertes et réutilisables. Les cadres législatif et administratif de l'open data ont été récemment définis par plusieurs lois[1] et par le Plan d'action national 2015-2017[2].

Bien qu'elle soit un des éléments clefs des discours d'accompagnement de l'ouverture des données, la notion de réutilisation reste mal définie, aussi bien par les acteurs publics que par les entreprises du numérique ou les chercheurs, chacun investissant cette notion

[1] Notamment, loi Macron, loi NOTRe, loi Valter et loi Lemaire.

[2] http://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/pgo_plan_action_france_2015-2017_fr.pdf. Le deuxième Plan d'action de la France pour 2017-2019 est en cours d'élaboration.

d'objectifs et d'attentes différents[3]. Malgré un nombre important de jeux de données publiques rendues disponibles ces dernières années *via* les portails métropolitains[4], le nombre d'applications exploitant les données ouvertes reste assez limité, tout autant que celui des utilisateurs et celui des services car ils n'atteignent pas des seuils de viabilité[5].

Cependant, il existe encore peu d'études empiriques sur la réutilisation des données ouvertes[6]. Parmi les publications récentes, nous pouvons citer le rapport *Re-using Open Data*, réalisé en 2017 par la Direction générale des réseaux de communication, du contenu et de la technologie de la Commission européenne[7]. En analysant les usages des données ouvertes par différentes organisations privées, ce rapport souligne l'inadéquation entre l'offre de données et les jeux de données les plus utilisés. Le document appelle les administrations publiques à mieux aligner l'offre et invite les entreprises à davantage communiquer autour des réutilisations réussies des données ouvertes.

Les deux enquêtes de terrain conduites dans le cadre de cette ANR ont cherché à comprendre le contexte et les pratiques de réutilisation des données ouvertes. La première enquête menée de septembre 2015 à juin 2016 se focalisait sur des réutilisateurs des données liées à la mobilité du portail des données métropolitaines data.grandlyon.com. Les résultats de cette enquête ont été exposés dans le livrable 1.1[8], consacré aux pratiques professionnelles, usages et besoins des acteurs réutilisant ce type particulier des données. La deuxième enquête de terrain, menée de février à avril 2017, analyse le cadre d'usage des

---

[3] 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*, v. 5, p. 117-132. URL: http://ojs.imodev.org/index.php/RIGO/article/view/204/338.

[4] Paquienséguy, F., Dymytrova, V. (2017). *Livrable n° 1.2 Analyse de portails métropolitains de données ouvertes à l'échelle internationale*. [Rapport de recherche] 1.2, Equipe d'accueil lyonnaise en Sciences de l'information et de la communication. 〈hal-01449348〉.

[5] Turki, S., Foulonneau, M. (2015). « Valorisation des données ouvertes : acteurs, enjeux et modèles d'affaires ». In : *Big data - Open data: Quelles valeurs? Quels enjeux?*, E. Broudoux, G. Chartron (Eds.), Louvain-la-Neuve, De Boeck Supérieur, p. 113-125.

[6] Kitchin, R. (2014). *The Data Revolution: Big data, Open data, data infrastructures and their consequences*, London, Sage.

[7] Berends, J. & al. (2017). *Re-using Open Data: a study on companies transforming Open Data into economic and societal values*. European Union. URL: https://www.europeandataportal.eu/sites/.../re-using_open_data.pdf.

[8] Paquienséguy, F., Larroche, V., Peyrelong, M-F., Vila-Raimondi, M., Dymytrova, V. (2016). *Synthèse des résultats de l'enquête auprès des ré-utilisateurs de données ouvertes : livrable n°1*. [Rapport de recherche] Sciences Po Lyon; Enssib; Lyon3. 〈hal-01432124〉.

données ouvertes par des développeurs, des data scientists et des data journalistes à l'échelle nationale, nous en présentons ci-après les résultats.

2. MÉTHODOLOGIE

2.1. NOTION DE CADRE D'USAGE

Du point de vue théorique, notre enquête s'inscrit dans les recherches consacrées à la dimension sociale de l'innovation technique qui portent une attention particulière aux interactions des divers acteurs de l'innovation[9] et, plus particulièrement, dans une approche centrée sur l'usager, ses difficultés, besoins et attentes[10] puisque l'objectif final d'Elico est de proposer des scénarios d'usages à des fins de réutilisations accrues des données ouvertes.

La notion de cadre d'usage qui est au centre de ce livrable vient des travaux de Patrice Flichy ; il distingue un « cadre de fonctionnement », qui renvoie aux fonctionnalités de l'objet et à l'usage technique, et un « cadre d'usage », qui se réfère à l'usage social. Les deux cadres ne sont pas fixes et s'élaborent au cours de processus d'ajustements complexes[11]. Leur articulation constitue le cadre socio-technique, qui « *reprend les histoires parallèles des mondes sociaux concernés* »[12]. Dans le cadre socio-technique de la réutilisation des données, chaque réutilisateur est porteur d'une histoire parallèle, liée à son propre écosystème et à ses propres finalités que nous tentons de saisir par notre enquête de terrain. Pour ce faire, nous avons décliné le cadre d'usage en trois catégories que nous avons tentées de saisir à

---

[9] Akrich, M. (1993a). « Les formes de la médiation technique », *Réseaux,* 60, 87-98 ; Akrich, M. (1993b). « Les objets techniques et leurs utilisateurs. De la conception à l'action », *Raisons Pratiques*, 4 : 35-57 ; Flichy, P. (1995). « L'action dans un cadre sociotechnique. Comment articuler technique et usage dans une même analyse? ». In : *Les autoroutes de l'information, un produit de la convergence*, J.-G. Lacroix et G. Tremblay (Eds.), Sainte-Foy, Presses de l'Université du Québec, p. 405-433 ; Flichy, P. (1995). *L'innovation technique. Récents développements en sciences sociales vers une nouvelle théorie de l'innovation*, Paris, La Découverte ; Flichy, P. (2008). « Technique, usage et représentations », *Réseaux*, vol.2, n° 148-149), p. 147-174. DOI : 10.3917/res.148.0147. URL: https://www.cairn.info/revue-reseaux1-2008-2-page-147.htm; Flichy, P. (2013). « Rendre visible l'information. Une analyse sociotechnique du traitement des données », *Réseaux*, vol. 2, n° 178-179, p. 55-89.

[10] Norman, D. A., Draper, S.W. (1986). *User Centered System Design: New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc. Hillsdale, NJ.

[11] Larroche, V., Dymytrova, V. (2017). « Le web sémantique, un moyen de visibilité et d'exploitation des open data auprès des communautés de réutilisateurs professionnels ? », colloque international pluridisciplinaire *Big data et visibilité en ligne : un enjeu pluridisciplinaire de l'économie numérique*, Université des Antilles, Fort-de-France, Martinique, 6-8 novembre 2017.

[12] Flichy, P. (1995). *Op. cit*., p.224.

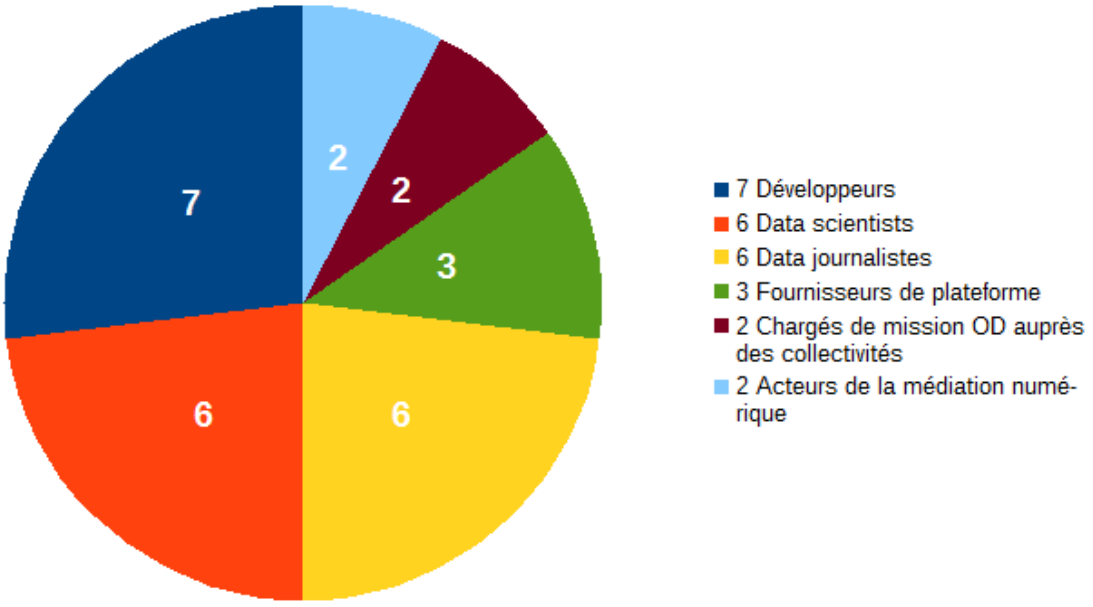
travers notre terrain : sources des données utilisées ; chaîne de traitement des données et outils et technologies professionnels mobilisés.

2.2. ENQUÊTE

Compte tenu de la diversité des profils, des finalités et, par conséquent, des pratiques de réutilisateurs des données, couplées à un objet d'étude très contemporain et en rapide progression, nous avons fait le choix d'une méthodologie qualitative[13]. Nous avons donc mené une série d'entretiens semi-directifs avec différents réutilisateurs professionnels des données en portant une attention au type de données utilisées, à la chaîne du traitement des données, aux outils et aux technologies utilisés, aux spécificités d'utilisation des données en temps réel et enfin aux différentes modalités d'utilisation des données (ANNEXE 1. GUIDE D'ENTRETIEN).

De février à avril 2017, V. Dymytrova et V. Larroche ont conduit 26 entretiens semi-directifs d'une durée moyenne de 45 minutes ce qui est une gageure avec des professionnels dont le temps est compté.

Figure 1. Répartition des interviewés par catégorie socio-professionnelle



Les professionnels interrogés ont été identifiés par deux moyens : certains noms ou entreprises nous ont été suggérés par les interviewés de l'enquête 2016, d'autres ont été

[13] Yin, R. K. (2009). Case *Study Research. Designs and Methods*, Newbury Park, Sage.

identifiés lors d'événements professionnels (salons, forums, conférences) ou encore *via* les sites web spécialisés (par ex., https://www.developpez.com/; https://www.lebigdata.fr). De fait, l'enquête a pris un caractère national, car nous avons interviewé des professionnels des données à Paris, à Lyon, à Toulouse, à Strasbourg, à Grenoble et à Brest.

Parmi les personnes interrogées, nous comptons 7 développeurs ; 6 data scientists/analystes, 6 data journalistes, 3 fournisseurs de portails et 4 personnes-ressources le chargé de mission Développement Numérique à la Région Rhône-Alpes-Auvergne ; le chef de projet Opendata de la Métropole de Grenoble ; le chargé de mission Innovation numérique au Tubà et les fondateurs de la coopérative Datactivi.st (*cf.* Figure 1).

2.2 CATÉGORIES D'ANALYSE DES ENTRETIENS

Tous les entretiens ont été enregistrés et ont fait l'objet de verbatim. L'analyse qualitative effectuée comprend les codages thématique et axial des propos des interviewés. Par rigueur scientifique, les entretiens ont été anonymisés, seuls le métier et le type d'organisation/d'entreprise seront ici mentionnés afin de rendre compte de la diversité des parcours et des situations des interviewés, issus de mondes socio-professionnels hétérogènes.

3. SPECIFICITÉS DES DONNÉES UTILISÉES

3.1. TYPES DE DONNÉES UTILISÉES

Les réutilisateurs professionnels ne se limitent pas à l'usage des données ouvertes car ils sont le plus souvent amenés à croiser plusieurs sources internes et externes de données. Les entretiens menés nous ont permis de spécifier ces sources et de mieux comprendre les conséquences du choix de la source sur la suite du travail des données.

Nos interviewés utilisent à titre professionnel cinq types de données[14] que nous détaillerons :

1. données ouvertes disponibles *via* des portails dédiés ;

[14] Nous attirons l'attention sur le fait que ces diverses données ne sont pas de même ordre. Les quatre premiers types renvoient aux données destinées à une réutilisation (données ouvertes, données collaboratives, données privées et données crowdsourcées) alors que le dernier type, données scrapées sont collectées à l'initiative des réutilisateurs à partir des sources publiques et privées.

2. données collaboratives/communautaires, créées et librement partagées par les utilisateurs *via* des librairies, catalogues ou plateformes ouvertes (de type *GitHub*, *OpenStreetMap*, etc.) ;
3. données privées fournies par des partenaires professionnels (clients, fournisseurs, intermédiaires, etc.) ;
4. données « crowdsourcées », générées par des usagers-consommateurs ;
5. données « scrapées » ou « crawlées »/collectées/compilées par des réutilisateurs eux-mêmes (*via* boîtiers, capteurs, robots d'indexation ou saisie manuelle).

Toutes ces données peuvent être statiques ou dynamiques. Elles renvoient à des données statistiques (par ex., données sociodémographiques), théoriques (par ex., cartographie, information voyageur) ou historiques (lorsqu'elles reflètent l'évolution temporelle des données). Les jeux de données s'accompagnent généralement de métadonnées qui font parfois l'objet d'un traitement spécifique. Comme le temps réel occupe une place particulière dans cette ANR, ces données en temps réel constituent pour nous un type particulier de données sur lequel nous reviendrons plus en détail à la page 17.

3.2. L'OPEN DATA

À l'heure actuelle, l'open data ne constitue donc pas la principale source des données pour les interviewés : « *…Aujourd'hui, l'open data, mais c'est peut être lié au fait que les projets sont encore peut être un petit peu verts, pas très matures dans les entreprises, on l'utilise dans 10-15% des projets, donc ça reste quand même une minorité des cas où on utilise de l'open data. Mais c'est une tendance qui a certainement vocation à se développer…* » (data scientist, fondateur et dirigeant d'un cabinet de conseil spécialiste de la donnée).

Un autre interviewé confirme ce constat : « *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*» (consultant formateur en data science).

L'open data intervient principalement en complément d'autres données aussi bien pour les développeurs que pour les data scientists: « *Ça permet de croiser les données que les entreprises ont, avec des informations qu'elles ne connaissent pas… Par exemple dans la prédiction, pouvoir croiser des données de trafic avec des données météorologiques, ça a un*

impact intéressant pour prédire des investissements sur les infrastructures qui sont en place » (data scientist dans un cabinet de conseil spécialiste de la donnée).

Plusieurs data journalistes interviewés soulignent que leur travail n'est pas construit autour des données ouvertes : « *Nous, on a souvent une manière de travailler qui va à contre-pied des données ouvertes puisqu'on essaye de compléter les données qu'on n'a pas, de trouver les données inédites qui n'existent pas ou qui ne sont pas publiques. Après, dans tous nos projets, on essaye de croiser ces données avec des données ouvertes* » (data journaliste co-fondateur d'une agence internationale spécialisée dans le journalisme de données).

Les données ouvertes le plus souvent citées par les professionnels interviewés proviennent de l'INSEE. C'est la base SIRENE (Système national d'identification et du répertoire des entreprises et de leurs établissements)[15], disponible en open data depuis janvier 2017, qui a été le plus souvent citée par les interviewés. Notons que la période de l'enquête (janvier-février) correspondait pour les professionnels des data à l'exploration de cette nouvelle source.

Malgré un fort taux de réutilisation, les jeux de données de ce répertoire ne permettent pas encore une géolocalisation précise des entreprises : « *Pour la base SIRENE, ce qui nous intéresse, c'est de pouvoir accéder à des données qui ont été ajoutées les trois derniers jours ou de récupérer les entreprises situées dans une zone géographique limitée. Il faut que l'API le permette. Etalab a mis en place un système qui permet de géolocaliser automatiquement les entreprises à partir des adresses mais les entreprises qui n'ont pas d'adresses précises sont mal géolocalisées (rue, commune...). Cela fait l'objet d'une collaboration avec Etalab et les collectivités, une sorte de crowdsourcing pour améliorer cela* » (développeur dans une société spécialisée en édition de logiciels).

[15] SIRENE rassemble des informations économiques et juridiques relatives à environ 10 millions d'entreprises et d'établissements, quel que soit leur secteur d'activité, situés en métropole ou dans les départements d'outre-mer. En moyenne, 10 000 modifications par jour sont enregistrées dans le répertoire. L'INSEE travaille avec de nombreux organismes, comme les greffiers des tribunaux de commerce, qui récoltent les informations sur les immatriculations, les radiations et les modifications au répertoire. Les données de la base SIRENE comprennent des données d'identification et des données économiques essentielles : l'adresse des établissements et leur statut juridique, le numéro SIRET/SIREN et le code APE. Source : https://www.etalab.gouv.fr/louverture-du-repertoire-sirene-par-linsee-au-1er-janvier-2017-une-avancee-majeure-pour-lopen-data.

Par ailleurs, à propos des données ouvertes, les réutilisateurs interviewés mentionnent deux autres thématiques phares : les données concernant des transports et la météo. Les data scientists s'en servent par exemple pour alimenter leurs modèles prédictifs concernant les usages des vélos en libre-service ou encore le remplissage des parkings. D'une manière générale, la donnée à propos de la météo est considérée comme ayant une forte influence sur les phénomènes socio-économiques : « *On en aura besoin pour prédire si on crée une réunion d'équipe à un moment ou à un autre parce qu'on a détecté que les réunions ont toujours eu lieu à un moment donné en hiver et à un autre moment en été, par exemple* » (lead data scientist dans une société de service en informatique spécialisée en logiciels libres).

Ainsi, remarquons nous que l'open data suscite à la fois beaucoup d'attentes par rapport à son potentiel d'exploitation et beaucoup de critiques par rapport à son indisponibilité et sa qualité chez les interviewés. Certaines questions liées à la réutilisation des données ouvertes ont déjà été soulevées dans le livrable 1, notamment celle de l'hétérogénéité des formats et des descriptions des données. Ce livrable n°3 recueille des témoignages et des avis inédits de réutilisateurs, data journalistes, data scientists et développeurs d'application et d'autres professionnels de la donnée ouverte travaillant à l'échelle nationale.

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 socio-professionnels sans une adaptation des jeux de données à ce nouveau contexte.

La présentation des données ouvertes se définit par rapport aux choix de leurs producteurs d'où une forte hétérogénéité des formats et des structurations : « *Si l'on veut avoir accès aux données sur les zones inondables sur une région, ces données sont gérées par des organismes départementaux qui utilisent des formats différents, il va falloir aller collecter les données dans des départements différents et faire des traitements pour consolider ces données au niveau de la région. Cela prend trop de temps. Du coup, ce n'est pas rentable »* (développeur dans une société spécialisée en édition de logiciels).

Les différents formats et présentations de données entre les villes constituent un frein dans la réalisation des applications et des cas d'usage portant sur plusieurs territoires : « *Faire une*

simulation à l'échelle d'une ville demande beaucoup de travail, à l'échelle de plusieurs villes il faut tout reprendre à partir de zéro, une plus grosse partie de travail doit être refaite, car il faut tout reprendre (les formats, les présentations ne sont pas les mêmes, les erreurs ne sont pas aux mêmes endroits) » (data scientist dans un établissement public de recherche).

Les interviewés souhaitent que les formats et la structuration des données ouvertes soient davantage homogénéisés : « *Ce serait parfait si chaque commune avait un jeu de données qui soit identique et surtout qui traite du même sujet* » (développeur, co-fondateur d'une start-up).

Un autre souhait exprimé par les professionnels interrogés concerne la constitution d'une base de données ouverte centralisée qui permettrait d'accéder à l'ensemble des données disponibles sur le territoire national. Selon eux, le portail national ne remplit pas pour l'instant ce rôle : « *Et datagouv.fr permet uniquement de savoir qu'il existe un tel ou tel type de données sur un territoire* » (développeur dans une société spécialisée en édition de logiciels).

Plusieurs interviewés souhaitent que les données ouvertes soient présentées dans des formats open source et dénoncent le recours à des solutions propriétaires : « *Les données des transports (la RATP par exemple) et la description des systèmes de transport public sont formatées dans un format qui a été imposé par Google, donc il y a des annuaires des villes qui fournissent ces données, qui sont publiques quand même, qui sont en OD mais ce qui est embêtant, c'est le fait que le standard est le standard privé pour l'intégration chez Google, il y a des démarches pas clean… Là, les enjeux sont non seulement techniques mais aussi éthiques* » (data scientist dans un établissement public de recherche).

Un autre point soulevé dans les entretiens concerne la mise à jour systématique des données ouvertes. Une fois publiés, les jeux de données en open data nécessitent une actualisation régulière pour être exploitables : « *Quand on donne des formations sur la data et on fait faire des exercices aux stagiaires avec des fichiers open data on va récupérer par exemple des fichiers sur le site de la mairie de Paris. Il y a des fichiers qui sont datés de 2008-2009 donc on sait que la fraîcheur de l'information dans le cadre d'une formation académique ou théorique ça va bien mais par contre dans le cadre d'une exploitation*

industrielle dans le monde de l'entreprise, elle est trop datée pour être pertinente » (data scientist dans un cabinet de conseil spécialiste de la donnée).

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. D'autres regrettent l'indisponibilité des données de l'Institut National de l'Information Géographique et Forestière ou l'absence des données concernant les aires de covoiturage à l'échelle nationale. Par ailleurs, la plupart des données ouvertes fournies en temps réel (transports, vélos) ne comportent pas d'historiques ce qui gêne le travail, notamment des data scientists qui utilisent des modèles prédictifs.

Les interviewés insistent sur le besoin d'informer davantage à propos de l'open data : «*Moi je baigne évidemment dans un environnement où les gens savent ce que c'est l'open data et l'utilise, professionnellement j'entends, mais par contre autour de moi, l'open data c'est pas forcément très parlant quand on se place du point de vue du simple citoyen qui n'est pas dans les métiers de l'informatique et de la donnée. Donc je pense que d'un point de vue de la communication il y a des choses à faire et des axes de progression* » (data scientist dans un cabinet de conseil spécialiste de la donnée).

Synthèse

D'une manière générale, les attentes des professionnels interrogés concernent :

- l'homogénéisation et la cohérence des formats et des structures de données similaires entre les différents territoires ;
- la présentation des jeux de données dans des formats open source ;
- la mise à jour systématique des données ;
- la mise à disposition de nouveaux jeux de données à propos d'autres thématiques ;
- la diffusion des informations à propos des données ouvertes.

3.3. DONNÉES COLLABORATIVES/COMMUNAUTAIRES

Comme nous l'avons dit en partie 3.1, les données ouvertes ne sont pas le seul type de données utilisées. Tous les professionnels interviewés utilisent les librairies et les référentiels ouverts pour rechercher les données qui leur manquent ou partager les données et les API (*Application Programming Interface*) qu'ils ont créées. La référence à *GitHub* est systématique dans les entretiens. Il s'agit d'une des plus grandes plateformes de développement de logiciels dans le monde qui compte 20 millions d'usagers et 57 millions référentiels de données en avril 2017[16]. Par exemple, un développeur interviewé y a trouvé des librairies qui lui ont permis d'accéder aux données du site web boncoin.fr qui n'a pas d'API publique.

OpenStreetMap[17] est pour nos interviewés une autre source importante des données collaboratives. Ce projet international fondé en 2004 a pour but de créer une carte libre du monde. Les contributeurs bénévoles du projet collectent et cartographient des données concernant les routes, voies ferrées, rivières, forêts, bâtiments, etc. Depuis 2012, ces données sont sous licence libre ODbl. Ce sont les développeurs qui recourent le plus souvent à *OpenStreetMap* pour récupérer par exemple les cartes, la typologie des rues, la Base Nationale d'Adresse Openstreetmap (BANO) ou encore les parkings : « *C'est une énorme base de données, pas très propre mais qui contient énormément de choses et donc ça permet d'avoir un premier jeu assez complet sur les parkings en France »* (directeur technique dans une société spécialisée en stationnement intelligent).

Selon plusieurs développeurs interrogés, *OpenStreetMap* demande plus d'effort au niveau de l'intégration des données que les API de *Google Map*, qui reste une solution propriétaire avec une excellente qualité du service, mais des quotas limitatifs pour un usage gratuit.

Enfin, une autre source de données collaboratives très sollicitée est *DBpedia*[18], un projet universitaire et communautaire d'exploration et d'extraction automatique de données de

[16] GitHub. (s.d.). The world's leading software development platform · GitHub: http://github.com.

[17] http://openstreetmap.fr/.

[18] Projet initié par l'université de Leipzig, l'université libre de Berlin et l'entreprise OpenLink Software : http://wiki.dbpedia.org/.

Wikipédia[19]. Ce sont surtout les data scientists qui l'évoquent : « *Pour l'analyse du langage, on va surfer sur DBpedia, qui est une base de données qui permet de faire une analyse sémantique qui fonctionne sur les articles de Wikipedia* » (lead data scientist dans une société de service en informatique spécialisée en logiciels libres).

3.4. DONNÉES PRIVÉES

Quant aux données privées, elles sont fournies aux réutilisateurs professionnels dans le cadre de contrats et de partenariats signés avec leurs producteurs ou détenteurs privés ou publics. Pour les data scientists et les data analystes, il s'agit d'une des principales sources de données : « *On travaille essentiellement pour des clients, donc on travaille beaucoup sur des données qui sont internes à nos clients, qui sont donc issues de leurs systèmes d'informations en propre. Ça peut être tous types de données liées à leurs activités, métiers. Ça peut être des données d'activités industrielles, d'activités marketing, d'activités de gestion des ressources humaines. Donc, c'est assez divers. Souvent on commence par traiter ces données internes là et dans un second temps, on est amené à les croiser avec des données externes* » (data scientist dans un cabinet de conseil spécialiste de la donnée).

L'accès aux données privées fait l'objet de négociations spécifiques ou peut être ponctuellement fourni dans le cadre de compétitions, de challenges ou d'hackathons. Pour continuer à utiliser ces données par la suite, les réutilisateurs doivent conclure des accords avec les producteurs : « *Le plus long ce n'est pas vraiment la technique, mais c'est entrer en contact avec les établissements et réussir à avoir accès à leurs données* » (développeur fondateur d'une start-up).

[19] Conçu par ses auteurs comme l'un des « noyaux du Web émergent de l'open data », *DBpedia* propose une version structurée des contenus encyclopédiques de chaque fiche encyclopédique de Wikipedia sous forme de données normalisées au web sémantique. *DBpedia* souhaite aussi relier à Wikipédia des ensembles d'autres données ouvertes provenant du Web des données. https://fr.wikipedia.org/wiki/DBpedia/.

3.5. DONNÉES « CROWDSOURCÉES » : L'EXEMPLE DE *MOOVIT*

Les professionnels interrogés intègrent aussi parfois dans leurs productions les données générées par les usagers de smartphones ou autres objets connectés (par ex., montre connectée dans le cas d'une application développée par la société *NextCairn* qui met en relation des sportifs et des entraîneurs[20]).

L'exemple de l'application *Moovit* est illustratif. La combinaison des données crowdsourcées et des données officielles des opérateurs constitue, selon le Country Manager de *Moovit*[21], que nous avons interviewé, le succès de l'application conçue pour faciliter les déplacements dans les transports en commun et réduire l'incertitude des usagers. Utilisée par plus de 70 millions de voyageurs dans plus de 1400 villes en 2016, *Moovit* a lancé son service dans 70% des villes grâce aux données cartographiées par les utilisateurs.

La start-up a ainsi réuni une grande communauté d'utilisateurs qui renseignent différents types de données, par exemple l'ajout des horaires, des arrêts sur une plateforme ouverte dédiée. Les usagers peuvent aussi partager leur localisation avec d'autres, donner un retour sur leur expérience des transports et aider les autres voyageurs. Dans ce cas, les usagers deviennent producteurs des données au profit de la start-up. Celle-ci revend ces données aux opérateurs qui les utilisent pour enrichir leurs propres données et produire des indicateurs de la qualité de service : « *Grâce au succès de l'appli parmi les usagers, on assiste à un véritable retournement de rapports de force, car aujourd'hui, on voit des opérateurs venir nous contacter parce qu'ils voient qu'on est capable de lancer un service qui a priori est meilleur que d'autres. Les données crowdsourcées constituent un argument dans les négociations avec les collectivités, Moovit promet un échange de données, dans le cas où les données des producteurs sont de mauvaise qualité* » (Country Manager de l'application pour la France).

L'usage des données crowdsourcées soulève ainsi plusieurs questionnements éthiques, notamment par rapport au degré de leur anonymisation : « *L'une des problématiques est*

[20] https://runreport.fr/.

[21] https://moovitapp.com/.

celle des données des usagers collectées par l'application, pour le moment chaque utilisateur a un compte et les données sont anonymisées dès le départ, mais si l'on veut devenir un fournisseur des données de masse, permettant le marketing ciblé, il faut trouver des solutions… » (développeur, co-fondateur d'une start-up).

3.6. DONNEES « SCRAPÉES » OU « CRAWLÉES »

À la différence des données ouvertes, des données collaboratives, des données privées et des données crowdsourcées, les données « scrapées » ou « crawlées » sont collectées à l'initiative des réutilisateurs, et ce par différents moyens. Il peut s'agir de mesures des phénomènes physiques effectuées par des boîtiers et des capteurs (par ex., les données de fréquentation des endroits publics) ou de la collecte de données sur le web et les réseaux sociaux grâce aux robots d'indexation et aux outils d'extraction de la donnée.

Les données captées peuvent provenir de capteurs existants, comme par exemple, dans le cas de l'application *Affluences*[22] qui informe en temps réel du taux d'occupation des bibliothèques municipales et universitaires. La plupart des établissements partenaires de la start-up en France disposent déjà des infrastructures générant les données sous forme de portiques antivol équipés des systèmes de comptage des passages. D'autres start-up fabriquent elles-mêmes les boîtiers connectés qui nourrissent leurs applications. Par exemple, grâce à ses propres boîtiers insérés au niveau des portes ou des barrières automatiques des parkings, *Copark*[23] récupère les données d'ouverture et de fermeture des barrières et génère ainsi les données sur le taux d'occupation. L'application met en relation les parkings ayant des places disponibles à certaines heures avec les automobilistes cherchant à se garer. Grâce à l'application, une place de stationnement peut être trouvée, réservée, payée et utilisée avec un smartphone. Quant à l'application *ParkingMap*[24], elle propose, dans certaines villes, d'intégrer les données issues du projet communautaire et collaboratif ParkingMapBox pour cartographier les places disponibles. Ce projet fait appel aux habitants pour héberger des capteurs sur leur rebord de fenêtre, visualisant ainsi l'état du stationnement en contrebas.

[22] http://www.affluences.com/.

[23] https://copark.co/.

[24] http://www.parkingmap.fr.

Si les données issues des capteurs sont souvent utilisées par les développeurs dans la production des applications, les données « scrapées » sont surtout récupérées et utilisées par des data scientists et des data journalistes. Les data scientists y recourent, par exemple, « *dans le cadre de la volonté de l'entreprise qui veut mieux connaitre ses clients ; il va être intéressant de croiser la connaissance qu'elle a en interne de ses clients avec la perception que les clients vont avoir de la marque, de l'entreprise sur les réseaux sociaux »* (data scientist dans un cabinet de conseil spécialiste de la donnée).

Les data journalistes mobilisent ces données pour la réalisation d'enquêtes sur un sujet spécifique : « *Je dois trouver moi-même un système pour récupérer les données qui sont publiques mais qui ne sont pas structurées et consolidées sous forme d'un fichier CSV ou Excel* » (data journaliste pour un site d'information spécialisé en actualité politique et juridique). Par exemple, pour étudier les dépassements d'honoraires par les médecins, un data journaliste interviewé a développé un script en PHP pour récupérer les données du site de la Sécurité sociale.

Dans le cas de collectes sur le web, la question des droits d'utilisation de ces données reste sans réponse : « *Si l'on fait simplement du scraping sur Internet, ce n'est pas de l'open data, on est sur l'acquisition des données dont on ne maîtrise pas forcément les droits d'usage* » (lead data scientist dans une société de service en informatique spécialisée en logiciels libres). « *Je ne me pose jamais la question, ce qui pas forcément bien, dans notre travail si les données sont disponibles, on va les publier, ce n'est pas correct, mais, on fait comme ça* » (data journaliste free-lance, ancien *du Monde* et de *Libération*).

Malgré l'automatisation croissante des collectes, les données manuellement saisies et compilées restent toujours importantes dans le travail journalistique : «*Par exemple pour une liste des soutiens d'un candidat à la primaire disponible sous forme d'un scan, j'ai dû saisir les données moi-même dans un tableur, c'est très fastidieux, j'ai dû faire ça plusieurs fois et au final, j'ai fini par utiliser un système de reconnaissance d'écriture, mais cela ne marche pas toujours bien* » (data journaliste pour un site d'information spécialisé en politiques publiques et actualités juridiques).

3.7. DONNÉES EN TEMPS RÉEL

Le temps réel occupe une place particulière dans cette ANR, consacrée au développement des solutions technologiques aidant à réutiliser les données issues de capteurs pour la création des applications liées aux déplacements intelligents.

Nos entretiens montrent une diversité de définitions du temps réel par différentes communautés socioprofessionnelles. En fait, la définition dépend du cadre d'usage des données. Dans tous les cas, plusieurs interviewés distinguent un « vrai/pur temps réel » d'un « temps réel effectif/en pratique » : « *La donnée temps réel pure, ce sera la donnée renvoyée par des capteurs des automates et qu'on peut obtenir à un rythme tendant vers l'infini. En pratique, il passe du temps entre la mesure effectuée par le capteur et le moment où il vous l'envoie, ce n'est pas en temps réel. Temps réel aujourd'hui c'est la donnée, idéalement à la seconde ou au moins toutes les minutes à chaque moment qu'on interroge le service »* (lead data scientist dans une société de service en informatique spécialisée en logiciels libres).

Le responsable d'un système d'information mobilité et déplacements métropolitain précise que la définition du temps réel dépend des contextes de production et d'usage des données : « *Le temps réel suppose la production de données incessantes, pour qu'elles soient disponibles au moment où l'utilisateur le souhaitera. Les contraintes des systèmes d'information aujourd'hui sont telles qu'on ne produit pas les données toutes les secondes et cela n'aurait pas forcément le sens. Par exemple, les données des bus à plusieurs niveaux (interurbains ou intra-urbains), il n'y a pas la même production. Pour les lignes intra-urbaines où il y a une grande fréquence de bus puisqu'on est au cœur de la ville, les données qu'on appelle temps réel sont produites toutes les 20 secondes (position et annonce du retard), pour les interurbaines, on va toucher aux populations plus rurales, on va remonter l'information toutes les 2 minutes… Cela a du sens en fonction des distances entre les deux arrêts* ».

Certains interviewés opposent le temps réel proposé par les portails OD aux informations fournies toutes les millisecondes : « *Je connais les gens qui travaillent dans la finance, là clairement, cela au-dessus d'une milliseconde, c'est du vrai temps réel. Dans le travail des traders, des brokers, cela va très vite… Dans le domaine de la mobilité, on a des données temps réel quasi immédiates, on a une évolution de la position d'un bus de l'ordre de 30 secondes, avec l'API de Cityway, par exemple, on a des infos toutes les 20-30 secondes sur*

une position. Pour l'avoir testé en vrai, cela ne marche pas bien. Il faut avoir une puce ou un GPS dans le bus pour pouvoir suivre un bus de manière quasi instantanée. Sinon ce n'est pas du vrai temps réel » (développeur free-lance open source).

Ces données intéressent surtout les développeurs qui proposent des applications fournissant aux usagers des informations en temps réel. Toutefois, certains ont besoin d'enrichir les données temps réel avec des historiques pour proposer des informations plus adaptées : « *L'algorithme tourne toutes les deux minutes et les données sont analysées et enrichies toutes les deux minutes avec les données prédictives. Le temps réel n'est pas forcément utile aux usagers, par exemple, en cas des infos sur le temps d'attente à la préfecture, le temps qu'on arrive sur place, la situation peut changer. Avec le temps réel et les historiques construites, nous faisons des analyses prédictives pour affiner les infos fournies aux usagers* » (développeur fondateur d'une start-up).

Pour les data scientists, il existe une continuité entre les données d'offre théorique[25], les données produites en temps réel et les données produites *a posteriori* (statistiques) : «*Quand on fait du prédictif, on veut pouvoir confronter ces prédictions au réel mais aussi au passé… Dans nos métiers, il y a vocation à construire un historique qui puisse ensuite servir de base pour des comparaisons analytiques entre le prédictif, ce que demain potentiellement on va faire, et ce qui s'est effectivement passé. Et ce sont des historiques qui peuvent remonter à une dizaine d'années parfois* » (data scientist dans un cabinet de conseil spécialiste de la donnée).

Plusieurs interviewés ont besoin de croiser les données en temps réel avec des historiques relatifs à ces données : « *Quand vous travaillez sur les données temps réel, vous avez besoin des données historiques, l'API de JCDecaux ne fournit pas d'historique, c'est pour ça, on garde les ordis allumés le jour et la nuit pour récupérer des historiques. Vous ne pouvez pas construire des modèles prédictifs si vous n'avez pas de données au moins sur une année, parce que je pense que les gens n'utilisent pas les Vélo'v de la même manière en janvier et en juin »* (lead data scientist dans une société de service en informatique spécialisée en logiciels libres).

[25] La notion d'offre théorique est bien définie dans le domaine des transports, où elle comprend la description topologique des lignes et du réseau, la description horaire et la description des services connexes. Voir http://www.territoires-ville.cerema.fr/referentiel-de-donnees-de-l-offre-de-transport-a1105.html.

Les data scientists qui utilisent les technologies du *machine learning* ont aussi besoin des historiques pour les données en temps réel : « *L'apprentissage automatique se fait sur les données historiques même si l'on a pour l'objectif d'analyser ensuite les données en temps réel* » (data scientist, fondateur d'une société spécialisée en valorisation des données des entreprises).

Les data journalistes sont ponctuellement amenés à travailler sur les données en temps réel, par exemple, pour informer en direct des résultats lors de l'élection présidentielle. Dans ce cas, le temps réel renvoie à des données mises à jour toutes les 15 minutes par le Ministère de l'Intérieur. Un autre data journaliste se réfère aux données en temps réel récupérées sur le portail Vigicrues[26] du Ministère de la Transition écologique et solidaire. Ici, le temps réel par rapport aux risques d'inondation renvoie à la fréquence qui peut varier de 10 à 15 relevés quotidiens. À l'instar de beaucoup d'autres données temps réel, ces données ne sont pas stockées sur le portail pour constituer des séries historiques. C'est pourquoi, afin de générer un article d'alerte sur les risques d'inondation, les data journalistes doivent constituer à partir du temps réel des historiques dans leur propre base de données pour analyser des évolutions sur des séries longues.

Plusieurs interviewés soulignent les difficultés face à l'implémentation des données en temps réel. Celles-ci concernent les manières de récupérer et de traiter les flux de données en continu mais aussi de les analyser : « *Le traitement de la donnée depuis quinze ans ou depuis vingt ans s'est fait de manière qu'on appelle cédulé, le scheduling, c'est le mot anglais pour planification, donc ça veut dire qu'on a un flux de donnée qui est planifié tous les jours ou toutes les semaines ou tous les mois pour faire un traitement de données. Et ça c'est le cas encore aujourd'hui classique dans les entreprises et qu'on sait très bien implémenter. Le temps réel, ça demande d'écouter la source de donnée, et dès qu'il y a un changement dans la source de donnée de capter ce changement. Techniquement c'est plus complexe et c'est plus consommateur en ressources informatique dans les serveurs, dans les machines… Typiquement pour les grosses entreprises qui ont des filiales dans plusieurs pays, il faut avoir une information à toute heure du jour et de la nuit par rapport à un fuseau horaire donné ou par rapport à un siège du groupe. Michelin, qui est un de nos clients a des filiales dans le*

[26] https://www.vigicrues.gouv.fr.

monde entier, donc pour que la donnée soit toujours fraîche, il faut récupérer de la donnée en temps réel sur tous les systèmes d'information du monde tout le temps en fait, ça ne s'arrête jamais. Donc, il y a des flux de données en continu et ce que ça signifie, c'est que les systèmes doivent inclure une notion de disponibilité, c'est-à-dire qu'ils ne doivent jamais s'arrêter de tourner» (data scientist dans un cabinet de conseil spécialiste de la donnée).

Malgré un grand intérêt des professionnels des datas envers le temps réel, ce type de données n'est pas systématiquement disponible en open data sur les portails métropolitains et quand il l'est, il s'accompagne rarement d'historiques en permettant l'usage par les réutilisateurs.

Synthèse

Dans cette sous-partie, nous avons présenté les différents types de données exploités par les professionnels, cependant chaque catégorie socioprofessionnelle a ses sources de données privilégiées en fonction des objectifs visés.

Les développeurs qui évoluent au sein des start-up s'intéressent beaucoup aux données en temps réel 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 malgré leur intérêt prononcé pour les solutions open source.

Les data scientists et les data analystes utilisent l'open data surtout pour former les étudiants, mais quand il s'agit de situations professionnelles, leur source privilégiée sont les données privées, fournies par des clients. Les données en temps réel intéressent les data scientists lorsqu'elles sont accompagnées d'historiques facilitant leur intégration dans les modèles prédictifs.

Enfin, les data journalistes sont très ingénieux dans les façons de rechercher des données sur différentes sources ; ils croisent divers types de jeux et de données. Ils recourent le plus souvent à des données collectées d'une manière journalistique auprès d'un réseau de contacts à des données crawlées ou scrapées sur le web par des robots d'indexation. D'une manière générale, les data scientists et les data journalistes effectuent un travail de veille sur les sujets liés aux données : « *Dans notre équipe aujourd'hui, on a une quinzaine*

de personnes, on demande à tout le monde de régulièrement d'ouvrir les yeux sur justement ce qui existe, ce qui est possible et ce qui se fait et d'enrichir nos espaces de capitalisation avec ces informations. On ne recherche pas forcément pour un projet de donnée, enfin si on a un projet donné dans lequel on sait qu'on va avoir de l'open data, on va le faire, mais en même temps on a en tache de fonds, entre guillemets, toujours un peu de récurrence pour garder les yeux grands ouverts sur ce qui se fait de manière à pouvoir ensuite être force de proposition » (data scientist, CEO dans un cabinet de conseil spécialiste de la donnée).

4. CHAÎNE DE TRAITEMENT DES DONNÉES

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

Au-delà des caractéristiques professionnelles, chaque réutilisateur de données est confronté à la chaîne de traitement comprenant la collecte et le stockage, l'exploration, la compréhension et l'analyse des données (enrichissement, croisement, recherche de corrélations, de classements), la transformation (traitement, nettoyage, alignement, annotation, indexation, etc.) et enfin, l'exploitation/implémentation des données : développement ou modélisation (application, data visualisation, cartographie, tableau de bord, article, etc.).

Guidée par des thématiques traitées, la recherche des données se fait le plus souvent par facettes et mots-clefs. Trois types de référencement facilitent le travail des professionnels :

- portails qui référencent les données au niveau national (par ex., datagouv, portails des collectivités) ;

- moteurs de recherche généralistes (par ex., Google, Yahoo) qui référencent beaucoup de données au niveau international ;

- enfin, les bibliothèques de données créées par des individus et partagées librement (par ex., GitHub). Ces bibliothèques sont bien visibles sur le web car elles respectent les formalismes des moteurs de recherche et des mots-clés.

Toutefois, retrouver une bonne donnée et savoir qu'elle existe peut toujours devenir un véritable défi, surtout elle relève d'un domaine spécifique.

Dans la suite du livrable, nous allons présenter les spécificités des chaînes de traitement des données et des outils professionnels des développeurs, des data scientists et des data journalistes.

4.1. DÉVELOPPEURS

Les développeurs produisent des applications web et mobile à destination des professionnels ou d'un large public citoyen. Ce travail se base sur une chaîne de traitement qui comprend les étapes suivantes (Fig. 2, page suivante) :

- récupération des données,
- intégration des données dans une base de données interne,
- transformation des données dans des formats compatibles avec les solutions utilisées,
- vérification des données,
- intégration des données dans une application web et/ou mobile.

Par exemple, pour produire *Proxiclic*[27] (Fig.3, ci-dessous), une solution de calcul et de représentation de l'accessibilité de services sur un territoire, destinée à des collectivités, la société *Datakode* utilise les données ouvertes issues de plusieurs sources : base de données SIRENE, base permanente des équipements et carroyage de l'INSEE. La chaîne du traitement de données se présente ici de façon suivante : la récupération des fichiers, leur intégration dans une base de données géographiques *OpenStreetMap*, la transformation de chaque

[27] http://www.datakode.fr/portfolio/proxiclic/.

carré (carroyage de l'INSEE) dans un point avec des coordonnées géographiques[28] et la mise en œuvre d'un web service permettant de faire des requêtes dans la base créée.

Figure 2. Chaîne de traitement des données par des développeurs.

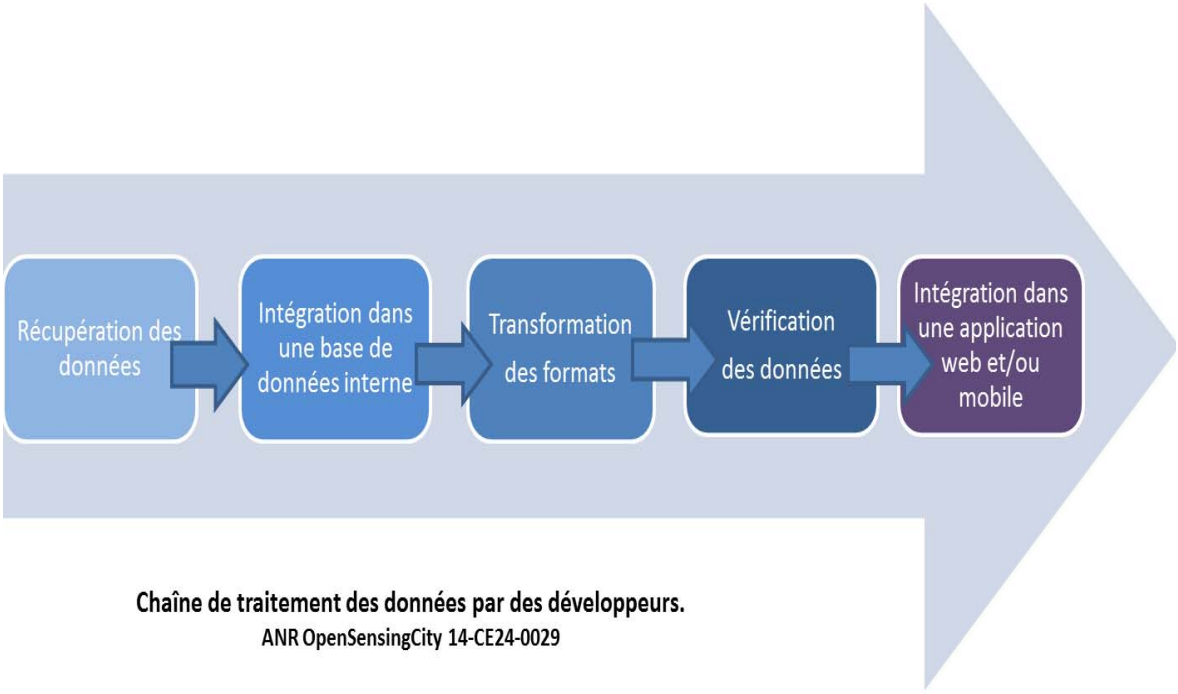
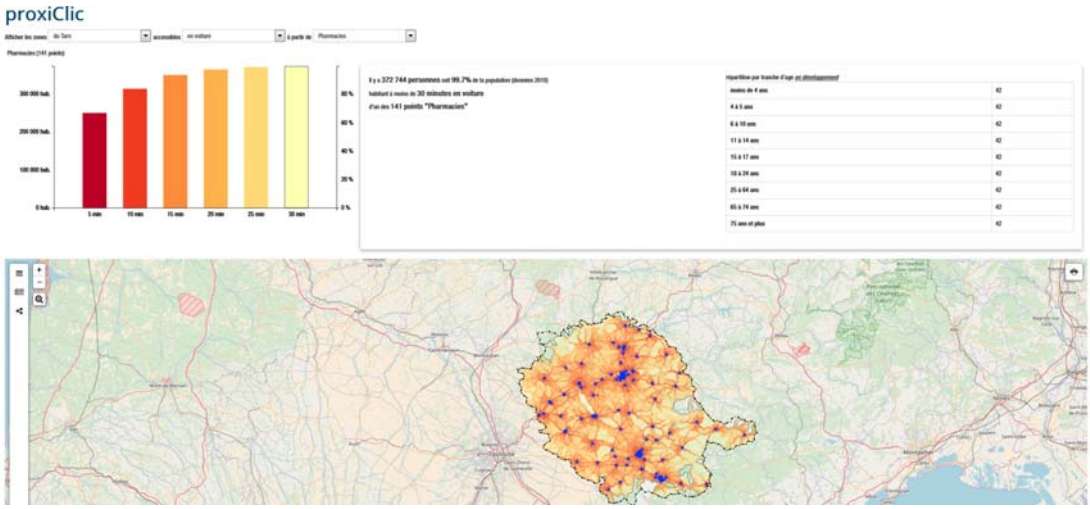


Figure 3. Capture d'écran de la solution Proxiclic : une visualisation des pharmacies accessibles en voiture dans le Tarn, 18 décembre 2017.



[28] La France est découpée en carrés de 200 mètres sur 200 m avec les informations sur le nombre d'habitants par carré. La transformation de chaque carré en point facilite l'obtention d'information sur une zone spécifique.

La réutilisation simple, facile et rapide des données collectées est souvent empêchée par les formats initiaux, en particulier quand il s'agit de formats propriétaires et par la structuration des jeux de données. Pour que les développeurs puissent interagir facilement avec les données en les recherchant et les récupérant automatiquement, ils ont besoin d'API, un ensemble de fonctions logicielles qui peuvent être appelés 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. De fait, pour l'exploration des données, certains développeurs utilisent des outils intégrés à Mozilla Firefox[29], qui permettent de « *se faire rapidement un avis sur n'importe quelle API, même si elle est cachée et non publique et de l'explorer d'une manière simple* » (développeur free-lance open source). Un autre outil utilisé par les développeurs pour tester une API est *Postman* de Google Chrome[30].

Figure 4. Solutions professionnelles des développeurs.

Développeurs	**Formats et langages :** JSON, CSV, GeoJson, Shape, KML, Python, GeoPy, SQL, Java Script, PHP, PHP Symfony 3 **Logiciels et API :** OpenStreetMap, Google Map, Postman de Google Chrome, FME, GeoServer, GeoNetwork, Csvkit, Addok

« *Au niveau de la chaîne de traitement, on a des systèmes en place sur notre serveur, ce sont des petites commandes qui sont lancées toutes les minutes ou cinq minutes ou quinze minutes etc. en fonction du parking qui derrière viennent automatiquement interroger les différentes URL pour récupérer l'information sur le stationnement. À chaque fois que l'on va avoir une nouvelle information, on va vérifier si cette information-là est différente avec ce qu'on a déjà récupéré. Si jamais la ville de Lyon nous dit que le stationnement change toutes les minutes, mais au final ça change toutes les cinq minutes, on ne va pas enregistrer cinq fois la même chose. Donc, déjà on a ce petit traitement... Et à partir de là, on a le résultat qu'on récupère, on va le convertir selon notre format standard qu'on a créé qui nous permet ensuite de l'enregistrer dans notre base de données… À partir du moment où on a récupéré nos données, ça met à jour la disponibilité du parking associé et tout de suite l'application*

[29] https://developer.mozilla.org/fr/docs/Web/API.

[30] https://chrome.google.com/webstore/detail/postman/fhbjgbiflinjbdggehcddcbncdddomop.

mobile va récupérer l'information. Donc là, c'est vraiment en une seconde, du moment où on récupère une nouvelle information l'application mobile va changer sa disponibilité, son code couleur ou voir en ce moment enlever le parking s'il est complet » (développeur et directeur technique d'une start-up).

Malgré leur présence, 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. Par exemple, lors du développement de l'application *Mobili.watch*[31], qui propose un système d'affichage personnalisé pour décompter le temps qu'il reste en fonction des déplacements en temps réel des transports en commun en Isère, un développeur nous a raconté comment il a dû faire face à une API défaillante : « *Il y avait une fonction dans l'API qui ne marchait pas et qui était essentielle pour notre idée. C'était l'accès aux données concernant les prochains passages des bus en temps réel pour pouvoir informer que le prochain bus arrive dans deux, trois ou cinq minutes. Sauf que cela ne marchait que pour environ 10% des arrêts de bus et de tramways de la région grenobloise. Donc, on a dû accéder à des données de deux manières : quand il y avait des données par ce moyen-là, on le prenait, sinon on bricolait autre chose, avec une fonction qui donnait des données moins temps réel, c'était moins intéressant pour nous, cela permettait que l'appli marchait mais on trichait… Comme c'est une petite société et une petite API, on a pu les contacter pour faire remonter l'information »* (développeur free-lance open source). Une autre difficulté à laquelle sont confrontés les développeurs vient des API privées qui fixent un taux d'accès limité, par exemple, lorsque le fournisseur de données limite en temps ou en nombre d'appels l'utilisation des ressources.

Plusieurs développeurs interviewés affirment que le format le plus couramment utilisé est actuellement JSON (JavaScript Object Notation – Notation Objet issue de JavaScript)[32] car il est facile à lire ou à écrire pour des humains et des machines : *« 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*

[31] https://mobili.watch/.

[32] https://www.json.org/json-fr.html.

que suffisamment simple pour que tout le monde travail avec » (développeur free-lance open source).

Si les données ouvertes sont récupérées au format CSV (l'un des formats traditionnels des portails métropolitains OD), les développeurs sont souvent obligés de les remettre dans un autre format, souvent le format standard d'encodage Unicode UTF-8[33]. 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 free-lance open source).

La palette des outils professionnels des développeurs interviewés est très variée et dépend des fonctionnalités et des thématiques proposées dans les applications développées. Nous remarquons cependant une préférence nette envers des logiciels open source dans notre panel. Elle peut s'expliquer, d'une part par le fait qu'il s'agit souvent de start-up en train de lancer leurs services et donc sans modèles économiques viables ; d'autre part, comme nous l'avons déjà évoqué, parce qu'il existe une proximité idéologique entre la communauté des réutilisateurs de l'open data et le mouvement open source.

Pour la gestion de leurs propres bases de données relationnelles, les développeurs utilisent le plus souvent des solutions de type SQL (Structured Query Language)[34] qui permettent de rechercher, d'ajouter, de modifier ou de supprimer facilement des données. Pour exploiter les bases de données géographiques, comme par exemple BAN (Base Adresse Nationale) et BANO[35], les développeurs utilisent le géocodeur Addock[36]. En même temps, plusieurs interviewés recourent au service de géocodage de l'API Google Map, permettant de récupérer les coordonnées géographiques d'une adresse ou d'une ville.

---

[33] https://fr.wikipedia.org/wiki/UTF-8.

[34] https://fr.wikipedia.org/wiki/Structured_Query_Language.

[35] La BAN (Base Adresse Nationale) est la base de référence nationale issue d'une convention signée entre l'IGN, le groupe La Poste, l'État et OpenStreetMap France. Elle fait partie des neuf bases de données de références, instituées par la loi pour une République Numérique du 7 octobre 2016, dite Lemaire.

[36] http://addok.readthedocs.io/fr/0.5.x-fr/.

Pour manipuler d'autres types de données géographiques, les développeurs utilisent GeoServer[37], serveur informatique qui permet de transformer la donnée dans différents formats géographiques (GeoJson, Shape, KML) à partir de la donnée source stockée dans une base de données ou sous forme de fichier. GeoNetwork[38] est une application de gestion des sources de données géo référencées. Elle édite et recherche des métadonnées et génère des visualisations interactives.

Pour automatiser certaines tâches, comme par exemple la récupération de données ou le développement d'un script ou d'un prototype, les développeurs recourent souvent au langage Python[39] et à des bibliothèques spécialisées qui lui sont associées.

Enfin, pour créer des sites web et des applications, les développeurs interviewés citent PHP (PHP Hypertext Preprocessor)[40] et ses différentes versions comme PHP Symfony.

Synthèse

Notre enquête a permis d'identifier un certain nombre de besoins et d'attentes des développeurs :

Importance des API et des web services comme modalités d'accès aux données ; beaucoup de jeux de données ouvertes sont encore publiés sous forme de fichiers lourds, qui nécessitent un téléchargement compliquant leur exploitation ;

Certitude quant à la disponibilité du jeu et des données dans la durée, sous le même format et aux mêmes conditions. Le changement de formats de données par le fournisseur intervient encore très souvent en cours de route, sans aucun effort d'information à destination des réutilisateurs ;

Accès à la documentation afin d'intégrer et d'interpréter correctement les données ;

Besoin d'une standardisation/homogénéisation des données et des métadonnées entre différents producteurs et territoires ;

[37] http://geoserver.org/.

[38] http://geonetwork-opensource.org/.

[39] https://www.python.org/.

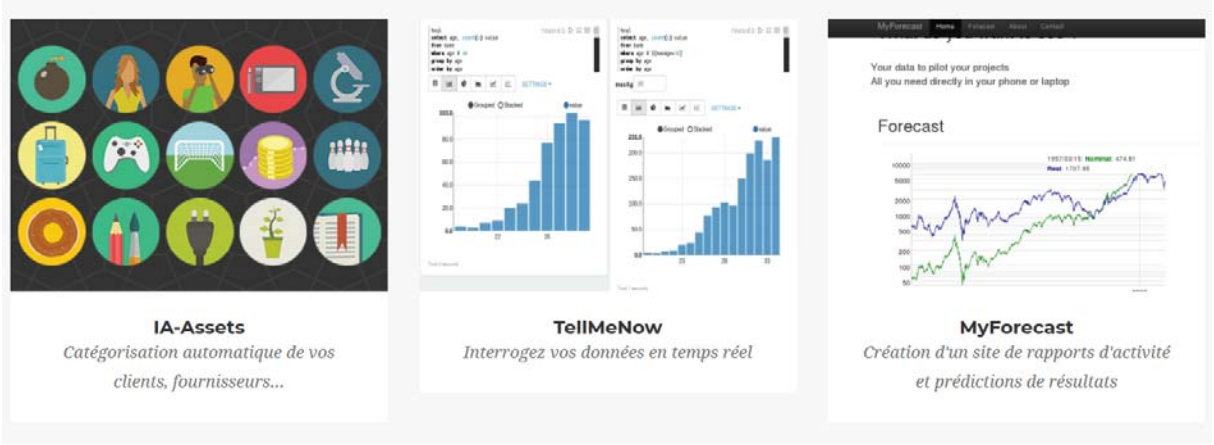
[40] http://php.net/manual/fr/intro-whatis.php.

Enfin, importance de l'utilisation des logiciels libres et open source par les établissements publics ; de nombreux réutilisateurs déplorent le recours aux solutions propriétaires.

4.2. DATA SCIENTISTS

Les data scientists mobilisent des modèles statistiques et mathématiques pour produire de l'information (*reporting*, tableaux de bord) avec une visée d'aide à la décision. En effet, ils doivent répondre précisément aux demandes, besoins et attentes des clients, qui sont principalement des grandes entreprises et organisations (Fig.5, ci-dessous). Pour cela, ils exploitent différents modèles (modèles d'optimisation, de simulation, de prédiction) et différentes méthodes (catégorisation automatique, analyse comportementale, ciblage).

Figure 5. Exemple des produits proposés par Dascils, société spécialisée en analyse des données (capture d'écran, 18 décembre 2017, URL : http://dascils.eu/).



La chaîne de traitement des données se présente pour les data scientists de la façon suivante :

- récupération des données,
- extraction/filtrage des données dont ils ont réellement besoin,
- transformation des données (nettoyage, structuration, modélisation, agrégation),
- intégration de données au sein d'un système interne ou d'un lac de données (*data lake*)[41],

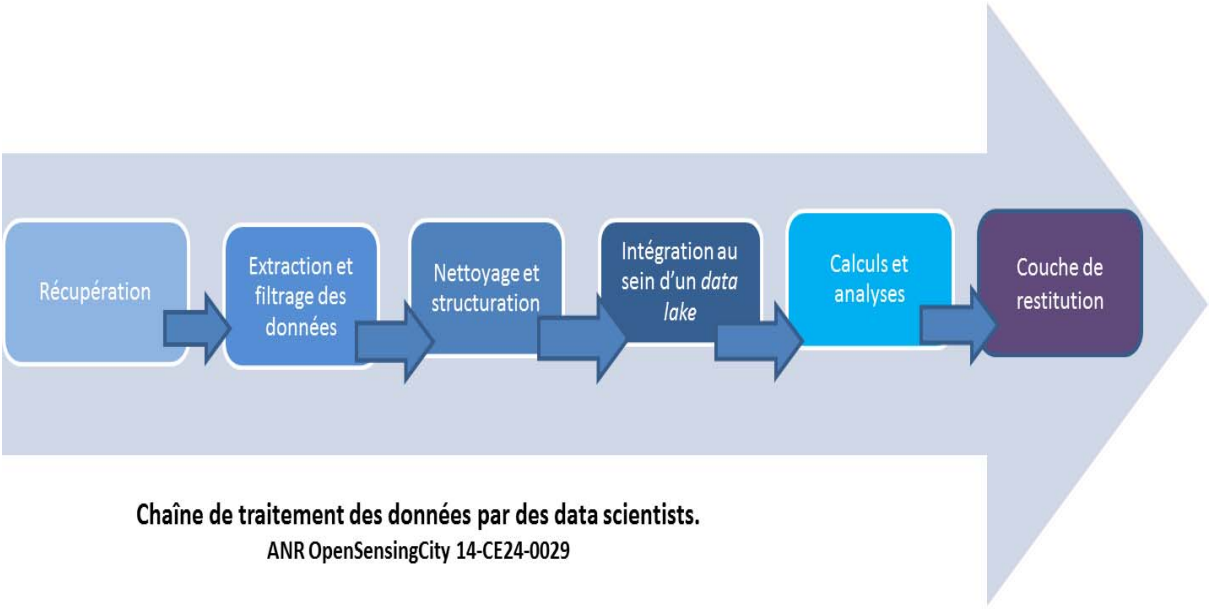
[41] Support de stockage universel pour des traitements complexes de grands volumes de données hétérogènes.

- calculs et analyses des données pour produire la couche de restitution qui prend le plus souvent la forme d'un tableau de bord (*reporting*) ou d'une visualisation (*dataviz*).

« Quand la donnée est dans une base de données déjà un peu structurée, c'est beaucoup plus simple, parfois, elle est vraiment cachée parce que vous voulez qu'une partie de la donnée et sous une certaine forme, et donc c'est là qu'arrivent les difficultés. Une fois la donnée récupérée et transformée, elle est intégrée au sein d'un data lake ou des outils élaborés par nous-mêmes, où l'on met les données dans un seul endroit et l'on peut faire les liens entre les données qui viennent de sources différentes et qui sont de types différents, par exemple une base de donnée, un fichier texte, un fichier qui vient d'Internet… Pour le stockage de ces données, on utilise des outils cloud mis à disposition par Microsoft et Amazon. Et après, on analyse les données, on a fait le choix de Power BI (Business Intelligence) de Microsoft pour la visualisation, cela permet d'avoir des rapports et des graphiques assez rapidement qui vont parler plus facilement aux clients » (data scientist, fondateur de la société spécialisée en valorisation des données des entreprises).

Les data scientists ont l'habitude de travailler avec des volumes de données importants appelés *big data*, ce qui rend nécessaire une phase de filtrage de valeurs et une phase de traitement mathématique ou statistique des données.

Figure 6. 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) demandent beaucoup de temps : « *Quand la donnée est dans une base de données déjà un peu structurée, c'est beaucoup plus simple, parfois, elle est vraiment cachée parce que vous voulez qu'une partie de la donnée et sous une certaine forme, et donc c'est là qu'arrivent les difficultés. J'ai un temps de code, effectivement. Il existe des outils qui simplifient cette méthode-là mais qui arrivent à une certaine saturation assez rapidement. Donc, c'est pour cela, pour moi, un data scientist est aujourd'hui, hélas, il est obligé de savoir coder pour pouvoir récupérer la donnée et la retransformer, sauf dans le cas des grands groupes qui peuvent se permettre d'avoir des data ingénieurs qui vont faire cette transformation* à leur place » (data scientist, fondateur de la société spécialisée en valorisation des données des entreprises).

Les data scientists sont bien outillés à toutes les étapes de la chaîne de traitement des données. En effet, pour l'extraction de la donnée, ils utilisent les logiciels de la famille ETL (Extract transform load)[42] qui transforment les données et font des opérations sur les données. Parmi ces logiciels, plusieurs interviewés citent une solution française open source : Talend[43]. Elle couvre un grand nombre des besoins en intégration et transformation des données.

Les outils du traitement utilisés dépendent du niveau de la structuration de la donnée: « *On ne traite pas toutes ces données de la même marnière évidemment. Pour les données structurées et semi structurées*[44]*, on va utiliser les ETL, qui embarquent des connecteurs qui permettent de se connecter à ces sources structurées et semi-structurées, donc là on va pouvoir nativement venir lire une base de donnée, lire un fichier Excel, etc. et récupérer*

[42] Extract-transform-load est connu sous le terme ETL, ou extracto-chargeur, (ou parfois : datapumping). Il s'agit d'une technologie informatique intergicielle (*middleware*) permettant d'effectuer des synchronisations massives d'information d'une source de données (le plus souvent une base de données) vers une autre https://fr.wikipedia.org/wiki/Extract-transform-load.

[43] https://fr.talend.com/.

[44] Les données structurées recouvrent les tableaux, les listes, les graphes… Elles sont gérées par des systèmes de fichiers ou de bases de données structurées, comme les systèmes relationnels (Système de Gestion de Bases de Données) et SQL (Structured Query Language). Les données non-structurées concernent les textes, les images…, leur traitement nécessite des opérations préalables pour en isoler les mots et les caractéristiques visuelles, les ramenant à des représentations semi-structurées.

l'information. Pour les données complètement non structurées, on va les aspirer par exemple en format brut et puis après, on va mener de l'analyse point à point. Par exemple pour les images ou pour les vidéos, on va travailler sur le pixel pour voir comment se situe l'information et on va retranscrire l'information de la donnée qui est exploitable. Donc là après, ce sont des processus qui sont évidemment plus longs et plus complexes quand la donnée est déstructurée » (data scientiste, fondateur et dirigeant d'une société spécialisée en analyse de la donnée).

Les *data lake*[45] jouent un rôle primordial dans le travail de data scientists : « *On va centraliser les données dans des systèmes qu'on appelle lac de données, donc un data lake 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 dans un cabinet de conseil spécialiste de la donnée).

Pour le stockage des données, les data scientists mobilisent les technologies de la fondation américaine *Apache*[46], qui développe beaucoup de projets open source. Plusieurs interviewés citent le socle d'application Hadoop[47] pour construire et modéliser les différents entrepôts de données. Le principe d'Hadoop est celui d'un traitement distribué de *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[48]. En effet, les outils particulièrement appréciés par les data

[45] Il s'agit d'un référentiel de stockage qui conserve une grande quantité de données brutes dans leur format natif jusqu'à ce qu'elles soient nécessaires : http://www.lemagit.fr/definition/Datalake-lac-de-donnees; http://www.journaldunet.com/solutions/dsi/1165409-qu-est-ce-que-le-datalake-le-nouveau-concept-big-data-en-vogue/.

[46] https://httpd.apache.org/.

[47] Hadoop est un framework libre et open source écrit en Java destiné à faciliter la création d'applications distribuées (au niveau du stockage des données et de leur traitement) et échelonnables (scalables) permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données. Ainsi chaque nœud est constitué de machines standards regroupées en grappes. Tous les modules de Hadoop sont conçus dans l'idée fondamentale que les pannes matérielles sont fréquentes et qu'en conséquence elles doivent être gérées automatiquement par le framework (https://fr.wikipedia.org/wiki/Hadoop).

[48] La scalability ou scalabilité (calque de traduction) désigne la capacité d'un produit à s'adapter à un changement d'ordre de grandeur de la demande (montée en charge), en particulier sa capacité à maintenir ses fonctionnalités et ses performances en cas de forte demande. https://fr.wikipedia.org/wiki/Scalability.

scientists interviewés 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 socle d'application Spark[49] en fait partie : *« Spark a aujourd'hui un peu le vent en poupe, en ce moment on en entend beaucoup parler. La particularité de ces systèmes, c'est que ce ne sont pas des bases de données traditionnelles, ce sont des systèmes qui stockent la donnée sous forme de fichier. Ça offre des possibilités de scalabilité très importantes puisqu'on va avoir x nœuds sur lesquels la donnée va être parallélisée, et si on a plus de volumes de données, on va rajouter des nœuds dans notre cluster. Plus on a de données, plus on rajoute et progressivement on est extrêmement scalable, ce qui n'était pas le cas des anciens systèmes de bases de données…* » (data scientist dans un cabinet de conseil spécialiste de la donnée).

Figure 7. Solutions professionnelles des data scientists.

Data scientists	**Formats et langages :** XML, Scala, Java, Apache Parquet, langage et bibliothèques Python **Logiciels et API :** R, QJIS, Hadoop, Power BI (Buisiness Intelligence) de Microsoft, Cplex d'IBM, Talend, MongoDB, Elastic Search, Dataiku, Scope, Spark, Qlik, Tableau

Le logiciel de traitement statistique R[50] est un autre outil incontournable des data scientists pour comprendre et analyser les données : « *Et à côté de ça, il y a un volet plus exploratoire qui est un volet data science entre guillemets, qu'on adresse aujourd'hui avec des langages comme R ou comme Python et qui consiste à aller croiser toutes les données qu'on a injectées : nos fameuses données de réseaux sociaux, nos données externes, nos données internes, nos open data. Et susciter des logiques et des croisements qu'on n'avait pas*

[49] Spark (ou Apache Spark) est un framework open source de calcul distribué (écrit en Scala, Java, Python et R) Il s'agit d'un ensemble d'outils et de composants logiciels structurés selon une architecture définie. Développé à l'université de Californie à Berkeley par AMPLab, Spark est aujourd'hui un projet de la fondation Apache. Ce produit est un cadre applicatif de traitements big data pour effectuer des analyses complexes à grande échelle. https://fr.wikipedia.org/wiki/Apache_Spark.

[50] R fonctionne sous la forme d'un interpréteur de commandes. Il dispose d'une bibliothèque très large de fonctions statistiques, d'autant plus large qu'il est possible d'en intégrer de nouvelles par le système des "packages", des modules externes compilés (sous forme de DLL sous Windows) que l'on peut télécharger gratuitement. R propose également une palette étendue de fonctionnalités graphiques. https://eric.univ-lyon2.fr/~ricco/cours/cours_programmation_R.html.

forcément envisagés en amont. Donc on parle de logique exploratoire parce qu'on ne sait pas forcément ce qu'on recherche avant de le trouver » (data scientiste, fondateur et dirigeant d'une société spécialisée en analyse de la donnée).

La restitution des résultats des analyses des données à destination des clients se présente principalement sous deux aspects : tableau de bord (*reporting*) et visualisation des données (*dataviz*). De fait, les data scientists doivent avoir des outils qui construisent et créent des restitutions assez ergonomiques avec des widgets plus lisibles et compréhensibles pour leurs interlocuteurs comme par exemple des histogrammes, des camemberts ou des cartographies. Pour cette étape, les solutions utilisées sont très variées, comme Qlik[51] et Tableau[52], édités par des sociétés américaines et *Power BI : Quick Insights* de Microsoft[53]. Les trois solutions citées comprennent plusieurs produits destinés à la création d'interfaces et de visuels soignés.

Synthèse

D'une manière générale, tous les data scientists interviewés ont souligné l'importance du travail de qualification des données : « *Une donnée n'amène de la valeur que si elle est qualifiée et qu'on a une qualité de donnée. Donc, si par exemple, on a une open data qui a été mal qualifiée et qui contient peu d'information ou qu'il y a des manques ou des trous dans l'information, elle peut ne pas amener finalement de valeur ajoutée. Dans tous les cas, quelles que soient les sources de données ; que ce soit une source de donnée interne, ou que ce soit une source de donnée de réseaux sociaux ou que ce soit des données ouvertes, nous on a toujours un travail de qualification pour valider le fait qu'une donnée amène de la valeur dans la chaîne data* » (data scientist dans un cabinet de conseil spécialiste de la donnée).

Notre enquête a ainsi permis de cerner les trois besoins principaux des data scientists dans leur travail de valorisation des données :

Avoir des données bien documentées, avec le maximum d'informations possible sur le

[51] https://www.qlik.com/fr-fr.

[52] https://www.tableau.com/fr-fr.

[53] https://docs.microsoft.com/en-us/power-bi/service-insight-types.

contexte de production, la structuration, la date de mesure, la mise à jour, la validité de la donnée et sur les marges d'erreur relatives à la donnée, notamment la précision de variance et la distribution de probabilité ;

Disposer d'une visibilité dans le temps sur l'évolution de la donnée et des unités de mesure utilisées pour la générer, car cela permettrait de produire les mêmes indicateurs dans la durée et les comparer entre eux.

Avoir des données sous un format compatible avec les outils qu'ils utilisent pour le traitement des données.

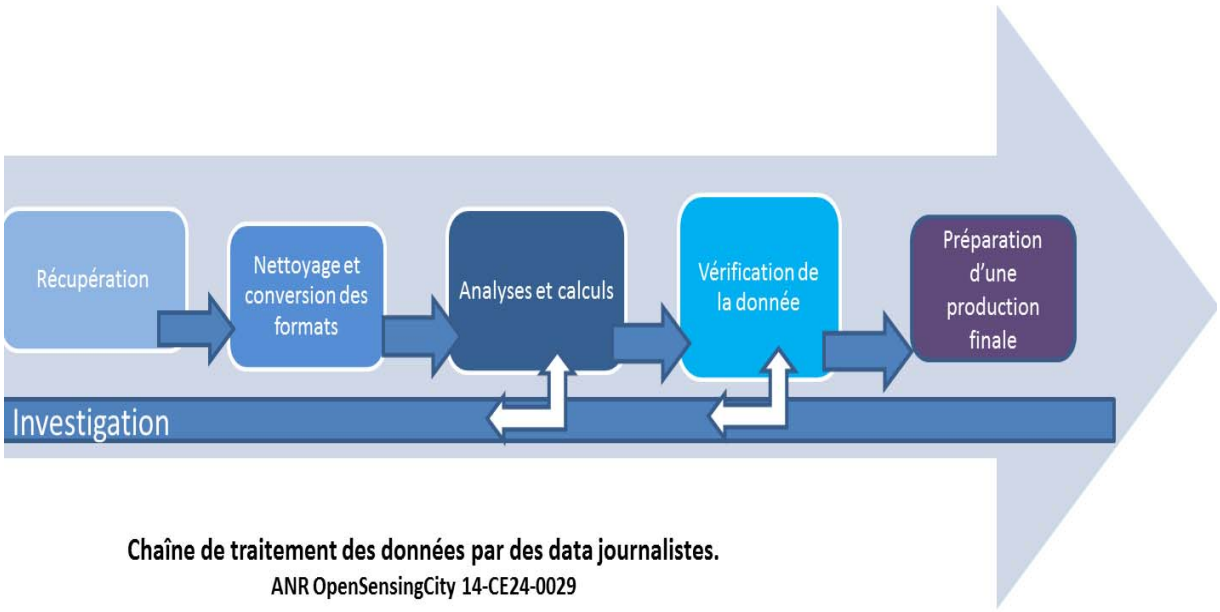
4.3. DATA JOURNALISTES

Les data journalistes utilisent les données pour mener des enquêtes, vérifier des informations (*fact-checking*) ou mettre en avant certaines données en apportant une plus-value éditoriale.

Figure 7. Capture d'écran de l'application réalisée par Rue89Lyon (18 décembre 2017, URL : http://www.rue89lyon.fr/2015/07/29/appli-region-lyon-ville-faite-pour-vous/).

Les productions finales prennent alors la forme d'articles de presse traditionnels et de mises en scène des données à travers des visualisations (*dataviz*), des cartographies, des graphiques, des comparateurs et des applications interactives (par ex. Fig. 7, p.34). La priorité des data journalistes est de se faire comprendre de l'ensemble de leur lectorat. Cela les oblige à passer beaucoup de temps sur les interfaces utilisateurs et sur le test des produits.

Figure 8. Chaîne de traitement des données par des data journalistes.



La chaîne de traitement de données des data journalistes (Fig.8., ci-dessus) comprend les étapes suivantes :

- obtention et récupération des données (extraction, intégration dans ses propres bases de données),
- nettoyage et conversion de formats,
- analyse des données pour les comprendre et trouver un angle d'attaque pour le sujet (calculs statistiques, croisement avec d'autres jeux de données, recherche de corrélations, visualisations intermédiaires),
- vérification de la donnée (confrontation avec d'autres sources, entretiens avec des spécialistes)
- préparation d'une production finale.

Toutes ces étapes s'accompagnent d'un travail d'investigation traditionnel mené en parallèle.

Parmi les data journalistes en France, il y a encore beaucoup d'autodidactes pour le versant data de leur activité professionnelle, les compétences, les formats et les outils mobilisés sont donc très variés. Pour extraire les données du web, les data journalistes interrogés utilisent *Outwit hub*[54] ou programment des scripts en Python. Ce langage open source est particulièrement apprécié par plusieurs interviewés car il s'appuie sur une grande quantité de bibliothèques, créées et maintenues par une large communauté d'usagers.

Les formats les plus utilisés sont Excel, XLS, JSON, CSV et KLS. La manipulation de différents formats ne pose pas de problèmes particuliers aux interviewés : « *Sur les formats, il n'y a pas trop de soucis. J'ai jamais été arrêté par un format ou un jeu de données, même si parfois on a dû multiplier par dix le temps de traitement* » (data journaliste pour le site web d'un quotidien régional).

La compréhension et le nettoyage des données constituent une étape qui peut s'avérer particulièrement chronophage : « *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 pigiste pour une rédaction locale d'un site d'information national).

Pour le nettoyage et la mise en forme des données, les data journalistes utilisent surtout le logiciel OpenRefine[55], anciennement Google Refine. Il permet de travailler les champs et de remédier notamment au problème très récurrent lié à l'encodage (présence/absence des caractères spéciaux, par exemple). Ce logiciel satisfait bien le besoin des journalistes d'avoir les 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*

[54] Logiciel conçu pour extraire et collecter automatiquement des informations à partir de ressources en ligne ou locales. Le programme reconnaît et récolte liens, images, documents, contacts, mots et groupes de mots récurrents, flux rss et convertit les données structurées ou non en tables formatées exportables vers des feuilles de calcul ou des bases de données (https://fr.wikipedia.org/wiki/OutWit_Hub).

[55] http://openrefine.org/.

propres, nettoyés où on a supprimé le niveau communal et on a tout mis au niveau cantonal. Dans la mesure où j'utilise les outils comme OpenRefine qui permettent d'identifier les erreurs et les modifier statistiquement, cela ne me dérange pas d'avoir les fichiers bruts, sales » (data journaliste free-lance, ancien *du Monde* et de *Libération*).

Figure 8. Solutions professionnelles des data journalistes.

Data journalistes	**Formats et langages :** Excel, XLS, JSON, CSV, KLS, JavaScript, Python **Logiciels et API :** OpenRefine, Google Spread Sheets, Google Fusion Tables, Infogram, ChartBlock, Gephi, Datawrapper, CARTO, QGIS, Mapbox, D3

Les data journalistes sont de grands utilisateurs de tableurs, logiciels d'édition et de présentation de tableaux. Certains travaillent avec Excel, d'autres lui préfèrent Google Spread Sheets[56] qui comporte des fonctionnalités similaires à Excel mais permet de partager les feuilles de calcul. Si les tableurs facilitent la manipulation des données, ils présentent aussi rapidement plusieurs limites qui obligent les data journalistes à se tourner vers le logiciel R pour des traitements statistiques plus complexes.

Pour des cartographies et des visualisations géospatiales (par ex., les cartographies électorales), les data journalistes utilisent la solution CARTO, ancien CartoDB[57], éditée par une société américaine ; le logiciel libre QGIS[58], porté par la Fondation Open Source Geospatial ; la solution Mapbox[59], proposée par une start-up américaine ; ou encore D3[60], une bibliothèque graphique en JavaScript. D'une manière générale, les bibliothèques logicielles sont très appréciées des data journalistes car elles fournissent des échantillons de codes et des exemples d'utilisation qui servent de base pour le développement des applications. Par exemple, pour réaliser une visualisation des marchés publics, un data

[56] https://docs.google.com/spreadsheets.

[57] https://carto.com/.

[58] https://www.qgis.org/fr/site/.

[59] https://www.mapbox.com.

[60] https://d3js.org/.

journaliste de *Rue89Strasbourg* s'est servi d'un exemple récupéré dans une bibliothèque en ligne, qu'il a développé et adapté par la suite[61].

« À partir du moment où je trouve le jeu de données, je l'importe, je le télécharge, je le nettoie, si on a une note méthodo, je commence par la note pour savoir ce qu'il y a dedans, ensuite je commence à jouer avec, en essayant de voir si par de simples tableaux croisés dynamiques, je peux mettre en avant quelques particularités ou si je peux le croiser avec d'autres données. Par exemple, j'ai une nouvelle donnée pour la délinquance, si je peux la croiser avec les données de la population d'une telle ou telle commune, je joue un peu avec pour voir s'il y a des éléments qui pourraient nous faire un angle du sujet. Ensuite, je vérifie s'il n'y a pas d'erreur statistique… Une fois, j'ai pu définir un angle d'article, je vais commencer à travailler cet angle-là, appeler les interlocuteurs pour nourrir l'article et travailler sur le meilleur mode de visualisation» (data journaliste pour le site web d'un quotidien régional).

Pour créer des infographies et des graphiques, les data journalistes interviewés citent : Google Fusion Tables[62], Infogram[63] et ChartBlocks[64] appréciés pour leur facilité de prise en mains. Gephi[65], logiciel libre d'analyse et de visualisation de réseaux fait aussi partie des outils mobilisés, notamment pour représenter le système de relations entre les acteurs.

Datawrapper[66] occupe une place particulière dans la panoplie des outils de visualisations car il a été spécialement développé pour ABZV, une organisation de formation des journalistes affiliée à BDVZ (Association des éditeurs de presse allemands) dans le cadre d'un programme de formation complet pour les journalistes.

Ce sont souvent les grandes rédactions qui se donnent les moyens de développeur leurs propres outils : « *Chez les Décodeurs du Monde, on a vu qu'on utilise toujours un type d'outils pour faire des cartes et du coup, on a créé un outil qui correspond aux usages de la rédaction,*

[61] http://www.rue89strasbourg.com/marches-publics-dou-sont-les-entreprises-attributaires-en-2014-84199

[62] https://support.google.com/fusiontables/answer/2571232.

[63] https://infogram.com/.

[64] http://www.chartblocks.com/.

[65] https://gephi.org/.

[66] https://www.datawrapper.de/.

cela génère automatiquement les cartes à partir des résultats électoraux » (data journaliste free-lance, ancien *du Monde* et de *Libération*).

Tous les data journalistes interviewés ont insisté sur le fait que les données ne suffisent pas pour produire une information et qu'elles ne remplacent pas l'enquête journalistique : « *Il ne faut pas penser que la vérité est à l'intérieur du tableau, il y a toujours un moment d'une enquête traditionnelle à faire pour corroborer ce que racontent les données »* (data journaliste pour un site d'information spécialisé en politiques publiques et actualités juridiques). Celle-ci se fait le plus souvent en parallèle avec l'analyse des données. À toutes les étapes de la manipulation des données, les data journalistes peuvent être amenés à mobiliser des personnes-ressources. Par exemple, ils contactent souvent les producteurs de données pour mieux comprendre les métadonnées : « *Moi, je lis rarement les métadonnées, où ils disent comment ils l'ont fait. C'est souvent très jargonné. Ce sont des choses qu'on va plus demander directement au ministère ou à la collectivité, pourquoi vous avez calculé cela comme ça* » (data journaliste pigiste pour une rédaction locale d'un site d'information national). Un carnet d'adresse d'experts est aussi indispensable pour vérifier et situer les résultats de l'analyse des données.

Les data journalistes ont une volonté forte de partager les données récupérées ou retravaillées. Ils le font systématiquement avec des confrères journalistes ou *via* un compte dédié sur *GitHub*. Par exemple, *Dataspot*, rubrique dédiée au data journalisme du *Télégramme de Brest* dispose d'un compte avec 400 fichiers sur le portail national *datagouv.fr*.

Synthèse

Notre enquête a permis de cerner plusieurs besoins des data journalistes concernant les données ouvertes :

- Avoir des données ouvertes sur différentes thématiques, y compris des données stratégiques ;
- Avoir des informations sur les données des administrations qui existent mais qui ne sont pas publiées en open data ;
- Disposer de données brutes pour être au plus près de la source ;

- Disposer de données ouvertes avec une forte granularité, régulièrement rafraîchies, dans une variété de formats et avec des différents modes d'accès (web services, API). Car les outils de traitement comme OpenRefine ne peuvent pas ouvrir et traiter des fichiers très volumineux ;
- Avoir des explications de différents champs et de nomenclatures particulières employées par les producteurs de l'open data pour bien en comprendre les données;
- Avoir une distance critique par rapport aux données car elles peuvent souvent comporter des erreurs.

5. SYNTHÈSE GÉNÉRALE

L'enquête de terrain réalisée montre que les contextes et les métiers liés à la réutlisation des données ouvertes sont aujourd'hui divers et variés. Les développeurs, les data scientists et les data journalistes sont des acteurs susceptibles de valoriser les données ouvertes auprès de plusieurs publics : habitants, usagers, entreprises et citoyens. Or, la plupart de nos interviewés utilisent les données ouvertes de manière ponctuelle, en les croisant toujours avec d'autres types de données. Pour avoir davantage de réutilisations pérennes de l'open data, il serait nécessaire de mieux prendre en compte les pratiques et les besoins des professionnels dans le secteur des data, avec toute la variété que nous venons d'exposer.

Malgré des objectifs et des productions différents, il existe des points communs dans la manière de travailler les données des développeurs, des data scientists et des data journalistes :

- Le traitement des données est une étape chronophage pour les trois catégories socioprofessionnelles, quels que soient leurs compétences et les outillages techniques mobilisés.
- La vérification et la qualification de la donnée est aussi très importante pour tous les professionnels interviewés.
- La présence des métadonnées et de la documentation expliquant les conditions de production des données est décisive pour interpréter correctement les données et proposer *in fine* des informations et des analyses fiables. Il est aussi important d'avoir des données structurées et décrites de la même manière afin de rendre interopérables les données analogues, mais issues de différents producteurs ou de différents territoires.
- Pour les interviewés, il apparaît important d'avoir des données disponibles dans une variété de formats, JSON et CSV restent actuellement les plus en vue. Les formats open source conviennent mieux à tous les interviewés car ils correspondent aux outils mobilisés mais également pour des raisons idéologiques, pour eux le mouvement de l'open data va de pair avec celui de l'open source.

- La panoplie des outils professionnels est très large. La démultiplication de l'ouverture des données par les administrations, les collectivités et les entreprises privées accroît le besoin d'outils de recherche et de navigation efficaces au sein de ces masses de données. Toutefois, l'ensemble des interviewés remarque la transversalité des usages des tableurs, du logiciel R, du langage de programmation Python et du SQL.
- Les réutilisations pérennes de l'open data (applications, visualisations) se basent souvent sur le recours à des API et à des web services comme modalités privilégiés d'accès aux données ; les systèmes d'information des producteurs de données doivent ainsi évoluer afin de résister aux problèmes de scalabilité.
- Enfin, pour nos interviewés, il existe un réel besoin d'informer davantage les entreprises et les citoyens de la démarche de l'open data et des jeux de données publiés.

ANNEXE 1. GUIDE D'ENTRETIEN

I. Données d'identification

- Entreprise, fonction, métier, domaines de compétence
- Parcours, formation

II. Productions réalisées à partir des données

- Quelles applications/quelles productions/quels articles/quelles visualisations/quels API/quels logiciels…
- Précisez la cible, les objectifs, le modèle économique

III. Chaîne de traitement des données

- Quelles données utilisez-vous ? Précisez les jeux de données, leur provenance, leurs formats, les licences liées à leur exploitation.
- Quelles sont vos pratiques de recherche des données ?
- Quel travail est opéré sur les données ? Si possible, décrivez la chaîne du traitement des données.
- Comment évaluez-vous la qualité de la donnée utilisée ?
- Quels descriptions/catégories/mots-clés associés à des jeux de données utilisez-vous ?

IV. Outils et technologies utilisées

- Quels sont vos outils professionnels ? Que vous permettent-ils de faire ?
- Quels sont vos besoins en termes d'outils facilitant la réutilisation des données ?

V. Temps réel

- Pour vous, qu'est-ce qu'une donnée en temps réel ?
- Quels sont vos besoins par rapport à ce type de données ?

- Quelles relations/combinaisons voyez-vous entre l'utilisation des données en temps réel et celle des données historiques ?

VI. Modalités d'utilisation des données

- Êtes-vous confronté aux questions éthiques liées à l'exploitation des données ?
- Voyez-vous un lien entre l'utilisation des données et la problématique des Smart Cities ?
- D'une manière générale, selon vous, quels sont les freins et les facilitateurs liés à l'utilisation des données ?

ANNEXE 2. COMPOSITION DU PANEL INTERROGÉ

Développeurs des applications (7)

- Développeur de l'application, fondateur et directeur d'une start-up (Paris)
- Développeur et directeur technique d'une start-up (Paris)
- Country Manager pour la France d'une start-up israélienne (Paris)
- Développeur et directeur technique d'une start-up (Lyon)
- Développeur dans une start-up (Lyon)
- Développeur free-lance open source (Grenoble)
- Développeur dans une société spécialisée en édition de logiciels (Toulouse)

Data scientists (6)

- Data scientist dans un établissement public de l'Etat à caractère industriel et commercial (Paris)
- Data scientist, consultant et formateur dans un centre de formation spécialisée en data science (Paris)
- Data scientist, fondateur de la société spécialisée en valorisation des données des entreprises (Lyon)
- Data scientist, fondateur et dirigeant d'un cabinet de conseil spécialisé en analyse de la donnée (Lyon)
- Lead data scientist dans une société de service en informatique spécialisée en logiciels libres (Lyon)
- Chargé de recherche en informatique dans un établissement public de recherche (Lyon)

Data journalistes (6)

- Data journaliste free-lance, ancien du Monde et de Libération (Paris)

- Data journaliste co-fondateur d'une agence internationale spécialisée dans le journalisme de données (Paris)
- Data journaliste pour un site d'information spécialisé en politiques publiques et actualités juridiques (Paris)
- Data journaliste dans une rédaction locale d'un site d'information national (Lyon)
- Data journaliste pigiste pour une rédaction locale d'un site d'information national (Strasbourg)
- Data journaliste pour le site web d'un quotidien régional (Brest)

Fournisseurs de portails (3)

- Société fournisseur de plateformes de l'open data clé en mains, à destination des organisations publiques et privées (Paris)
- Start-up, fournisseur de plateformes de mutualisation de données ouvertes des collectivités (Paris)
- Start-up, producteur d'un annuaire global de l'open data et des services associés (Lyon)

Personnes-ressources (4)

- C. Lambert, chargé de mission Développement Numérique à la Direction des Infrastructures et de l'Economie Digitale de la Région Rhône-Alpes-Auvergne (Lyon)
- F. Petit, chef de projet Opendata de la Métropole de Grenoble (Grenoble)
- B. Loeillet, chargé de mission Innovation numérique au Tubà (Lyon)
- J. Gombin, S. Goeta, fondateurs de la coopérative Datactivi.st (Paris)