HAL
open science

La fabrique des données brutes. Le travail en coulisses de l'open data

Jérôme Denis, Samuel Goëta

To cite this version:

Jérôme Denis, Samuel Goëta. La fabrique des données brutes. Le travail en coulisses de l'open data. Penser l'écosystème des données. Les enjeux scientifiques et politiques des données numériques, Feb 2013, Paris, France. halshs-00990771

HAL Id: halshs-00990771

https://shs.hal.science/halshs-00990771v1

Submitted on 14 May 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

La fabrique des données brutes

Le travail en coulisses de l'open data

Jérôme Denis
Telecom ParisTech
46, rue Barrault 75013 Paris - France
jerome.denis@telecom-paristech.fr

Samuel Goëta
Telecom ParisTech
46, rue Barrault 75013 Paris - France
samuel.goeta@telecom-paristech.fr

Journée d'études SACRED « Penser l'écosystème des données. Les enjeux scientifiques et politiques des données numériques » Paris - 13 février 2013

Résumé

Depuis quelques années, les initiatives d'open data se sont multipliées à travers le monde. Présentées jusque dans la presse grand public comme une ressource inexploitée, le « pétrole » sur lequel le monde serait assis, les données publiques sont devenues objet de toutes les attentions et leur ouverture porteuse de toutes les promesses, à la fois terreau d'un renouveau démocratique et moteur d'une innovation distribuée. Comme dans les sciences, qui ont connu un mouvement de focalisation similaire sur les données et leur partage, l'injonction à l'ouverture opère une certaine mise en invisibilité. Le vocabulaire de la « libération », de la « transparence » et plus encore celui de la « donnée brute » effacent toute trace des conditions de production des données, des contextes de leurs usages initiaux et pose leur universalité comme une évidence. Ce chapitre explore les coulisses de l'open data afin de retrouver les traces de cette production et d'en comprendre les spécificités. À partir d'une série d'entretiens ethnographiques dans diverses institutions, il décrit la fabrique des données brutes, dont l'ouverture ne se résume jamais à une mise à disponibilité immédiate, évidente et universelle. Il montre que trois aspects sont particulièrement sensibles dans le processus d'ouverture : l'identification, l'extraction et la « brutification » des données. Ces trois séries d'opérations donnent à voir l'épaisseur sociotechnique des données brutes dont la production mêle dimensions organisationnelles, politiques et techniques. Plutôt que d'écarter l'idée de données brutes en les qualifiant de mythe ou d'illusion, la mise en lumière de ces opérations invite à prendre au sérieux ce vocabulaire et comprendre que si les données ne sont jamais données au sens de « déjà là », ce travail permet en revanche d'en faire des données au sens de « don ».

« Yet in the field of science studies, we have in general focused attention on what scientists do with data, rather than on the mode of data production and storage. » (Bowker, 2000, p. 661)

L'open data et ses coulisses

Depuis quelques années, les programmes d'open data ont fait entrer les données brutes des administrations publiques en politique [1]. Présentées comme objets essentiels d'une transparence renouvelée, instruments d'une nouvelle *accountability* institutionnelle, voire ressources pour l'innovation, leur mise à disposition au plus grand nombre est devenue un enjeu crucial qui s'est cristallisé dans une multitude de projets à travers le monde, l'ouverture des données ayant été instaurée comme une obligation légale dans de nombreux pays[2]. On peut désormais trouver sur de nombreux portails Web des jeux de données extrêmement variés, diffusées par des États, des villes, des grandes institutions, et parfois des entreprises.

Les politiques d'open data participent au mouvement de transparence généralisée qui semble s'installer dans de nombreux domaines d'activité et qui nourrit ce que certains ont appelé des « cultures de l'audit » (Power, 1997; Strathern, 2000). Elles étendent le domaine de l'*accountability*, et peuvent être considérées comme l'un des outils phares de la gouvernementalité des sociétés techniques, tournée vers la figure du citoyen informé (Barry, 2001). Dans cette perspective, l'ouverture des données brutes des administrations et les dispositifs qui incitent à leur utilisation apparaissent comme des moteurs de la reconfiguration non seulement de l'État, qui se donne à voir sous un jour nouveau, mais aussi des citoyens, instaurés en public du gouvernement et de ses données (Ruppert, 2013).

En insistant sur ces aspects et en ne se focalisant que sur les conséquences politiques de l'instauration d'une transparence généralisée par la diffusion exponentielle de données, on risque toutefois de négliger un aspect important. À prendre pour acquis la circulation même des données, on laisse de côté les processus concrets qui amènent à leur ouverture même. Ce point est d'autant plus sensible que les programmes d'open data, tout comme les acteurs qui luttent pour leur généralisation, mobilisent un vocabulaire de l'ouverture et de la libération, partagent des métaphores empruntées au domaine des ressources naturelles, qui font de la donnée un allant de soi, un « pétrole », une entité naturellement disponible qu'il suffirait de libérer pour produire de la transparence et favoriser l'innovation. Mais que sait-on exactement

[1] La revendication pour la diffusion de données brutes a été formulée dès la rencontre de Sébastopol de 2007 qui a posé les bases de l'open data. Parmi les huit principes formulés, le second souligne l'importance d'accéder à des données sans traitement dites « primaires ». « Data Must Be Primary: Data are published as collected at the source, with the finest possible level of granularity, not in aggregate or modified forms. », in 8 Principles of Open Government Data. http://www.opengovdata.org/home/8principles

[2] Le G8 de Lough Erne qui s'est tenu en 2013 a donné lieu à la rédaction d'une charte qui stipule que l'ouverture des données sera la pratique par défaut des administrations, une obligation assortie de plans d'actions détaillées dans les pays signataires.

des conditions dans lesquelles les données sont ouvertes et diffusées ? Entre quelles « petites mains » (Denis & Pontille, 2012) les données passent-elles avant d'être libérées ? Sont-elles toujours déjà-là ? Suffit-il de les transférer d'un système d'information interne vers un portail en ligne pour mettre en œuvre une politique d'open data ?

Pour tenter de répondre à ces questions, il est utile d'observer ce qu'il s'est passé, quelques années auparavant, dans les domaines scientifiques au sein desquels la publication et l'échange de données sont devenus des enjeux cruciaux. Depuis une vingtaine d'années, nombreuses sont les disciplines au sein desquelles la collecte des données s'est peu à peu découplée de la production des résultats. La constitution de bases de données, tout comme la publication de données non traitées ont été ainsi hissées au statut de productions scientifiques à part entière (Bowker, 2000). Cette nouvelle place faite aux données s'est accompagnée de transformations majeures aussi bien dans l'organisation du travail scientifique (Edwards, 2010; Hine, 2006) que dans les résultats eux-mêmes, pris dans un « déluge » par lequel le processus même de la découverte scientifique s'est modifié, représentant pour certain un quatrième paradigme scientifique (Bell, Hey, & Szalay, 2009).

Si l'on commence à saisir aujourd'hui l'ampleur des conséquences du centrage d'une partie de l'activité scientifique sur les données, on sait mal évaluer ce que les politiques d'open data feront, et ont déjà fait, au sein des administrations. D'autant moins que la majorité des regards semblent aujourd'hui tournés en aval, du côté de l'évaluation des usages des données ouvertes, de leur impact sur l'innovation ou, comme nous venons de le voir de leurs conséquences politiques. Dans cet article nous proposons d'interroger ce mouvement d'ouverture des données publiques en nous inscrivant dans la suite des méthodes et des questions de recherche développées dans le cadre des *Science and Technology Studies*. Cela passe par une posture particulière face aux données publiées, qui consiste à privilégier une analyse des conditions concrètes de leur production et de leur circulation, au-delà des seules injonctions à l'ouverture. Une telle perspective permet de remettre en question la définition positiviste que charrie, tant en science que dans les politiques d'open data, le vocabulaire de la donnée brute (Gitelman, 2013). Plutôt que de faire des données un allant de soi, ingrédient transparent qu'il suffirait d'échanger pour équiper des travaux de recherche à l'échelle internationale, ou de libérer pour faire advenir une nouvelle transparence politique et stimuler dans le même temps l'innovation, nous défendons l'importance d'étudier au plus près les conditions concrètes de leur *fabrication*. Nous pensons en effet qu'il y a un intérêt à la fois théorique et pratique à mettre en lumière les enjeux concrets de la production des données ouvertes, au même titre que celles des sciences du climat par exemple (P. N. Edwards, 2010). À travers une exploration ethnographique des coulisses de l'open data[3], nous pouvons donner à voir l'épaisseur sociotechnique des données publiques, comprendre ce que leur ouverture implique à la fois sur les plans politiques, organisationnels et techniques, et de manière plus générale interroger les éventuelles spécificités des données publiques par rapport aux données qui ont fait l'objet des premiers travaux sur les infrastructures informationnelles en science (Edwards et al., 2009).

[3] Cette ethnographie rassemble plusieurs types de matériau, afin d'appréhender les multiples facettes du travail d'ouverture des données. Des entretiens approfondis ont été menés dans des collectivités locales, des institutions et une entreprise, auprès de personnes en charge de la mise en oeuvre du programme open data, de gestionnaires de données et de responsables informatique. La démarche a été complétée par l'observation de comités de pilotage de deux projets open data ainsi que l'observation participante d'un programme en cours de préparation. Enfin, une série de documents internes et externes (plaquettes, articles de presse, billets de blogs et commentaires) a été assemblées et analysée.

Dans un premier temps, nous proposons de revenir sur les principaux enseignements des *Science and Technology Studies* quant aux données, afin de définir plus précisément notre posture théorique et méthodologique. Nous nous arrêterons ensuite sur trois dimensions essentielles à la compréhension des processus d'ouverture de données publiques : l'identification, l'extraction et la « brutification ». Ces trois aspects montrent, sur des points différents, que les données publiques ne sont pas disponibles en l'état, prêtes à être libérées. Leur existence même est loin d'être une évidence. Les explorations menées par les responsables de l'open data pour identifier des données susceptibles d'être ouvertes, et les transformations dont elles font l'objet montrent que le caractère « brut » d'une donnée n'est pas une qualité intrinsèque, qui serait liée à sa primauté dans un circuit d'échange, mais le résultat d'opérations spécifiques.

Au fil de ces arguments, nous verrons ce que les *Science and Technology Studies*, et notamment les *Infrastructure Studies*, peuvent apporter à l'étude de l'open data, mais aussi ce que les pratiques d'ouverture des données dans les administrations peuvent apporter à la compréhension des infrastructures informationnelles contemporaines.

Le travail des données dans les *Science and Technology Studies*

Étudier les conditions concrètes de la production des connaissances est une posture classique en sociologie et en histoire des sciences. C'est évidemment le mouvement principal, radical, qu'ont entrepris les premières ethnographies de laboratoires (Knorr-Cetina, 1981; Latour & Woolgar, 1988; Lynch, 1985). En s'attachant à explorer méthodiquement les espaces de travail des scientifiques, ces travaux ont apporté une compréhension fine des opérations successives qui sont nécessaires à la production des résultats scientifiques. Ces opérations constituent des coulisses au sens de E. Goffman (1973) : leur accès est généralement réservé à quelques spécialistes et très peu d'entre elles sont rendues publiques. L'enjeu des ethnographies de laboratoires n'était pas de remettre en cause l'activité scientifique ni de la critiquer, contrairement à ce que certains ont pu comprendre, mais au contraire d'en souligner la richesse et la complexité.

Sans prétendre à l'exhaustivité, trois dimensions peuvent être retenues de ces très nombreux travaux, trois dimensions qui nous guiderons dans l'analyse des coulisses de l'ouverture des données publiques. La première est intrinsèque au geste même des premières ethnographies de laboratoire et de leur ancrage dans la sociologie du travail. En appréhendant les activités scientifiques en tant que pratiques professionnelles, ces travaux ont souligné leur caractère situé, organisé, mettant en lumière les nombreuses opérations — aussi bien manuelles qu'intellectuelles — qui les composent. Cette sociologie du travail scientifique a notamment permis de remettre en question deux figures clefs des versions officielles de la science : celle du génie individuel et celle de la découverte désincarnée.

Face à la figure du scientifique isolé, les enquêtes ont montré l'importance de l'organisation collective du travail scientifique, dont les tâches débordent largement le cadre des manipulations en laboratoire et de l'écriture des articles à proprement parler. Dans la tradition de la sociologie du travail initiée à Chicago, ses travaux ont pu montrer que les laboratoires scientifiques reposaient sur une division morale du travail par laquelle techniciens et petites mains sont en partie invisibilisés, notamment dans la production officielle des résultats (Barley & Bechky, 1994; Shapin, 1989). Au-delà de l'organisation du travail, c'est donc la définition de la contribution des uns et des autres qui se jouent dans la production scientifique (Pontille, 2013). Parce qu'elle est étroitement liée aux instruments et à leur manipulation, la question de

la division morale du travail a fait l'objet d'un intérêt renouvelé avec l'apparition des grandes bases de données scientifiques. Aux techniciens de laboratoires, aux doctorants et autres emplois précaires de la science, se sont ajoutés de nouveaux métiers aux contours et à la place encore incertains, comme les data managers (Baker & Millerand, 2009). De même, la prise en charge des métadonnées peut être considérée comme un « sale boulot » dont personne ne veut se charger (Edwards et al., 2011).

À la figure de la découverte désincarnée, les ethnographies de laboratoires ont également opposé une documentation précise et approfondie des nombreuses opérations concrètes qui font le quotidien des laboratoires et mènent des premières expérimentations jusqu'à la publication en passant par les différents recueils de données, leur traitement, mais aussi la recherche de financement, le montage de partenariat, etc. Sur ce versant, le travail scientifique apparaît beaucoup plus impur qu'on pourrait le croire de l'extérieur. Les ethnographies de laboratoires ont insisté sur les bricolages sur lesquels ces opérations reposaient, non pas pour les dénoncer, mais au contraire pour en montrer la complexité et la délicatesse. Cette perspective a notamment été développée pour mettre en lumière la chaîne des transformations mises en œuvre dans la « réduction » des données scientifique et la fabrication des résultats destinés à circuler après qu'ils aient été figés par la publication, devenus des « mobiles immuables » (Latour, 1985). Récemment, l'intérêt s'est aussi déplacé en amont, dans les activités de « récolte » de données elle même et les différentes manipulations dont elles font l'objet avant même le processus scientifique traditionnel. Ces enquêtes ont mené à la remise en cause de la notion de données brutes (Gitelman, 2013) sur laquelle nous reviendrons, mais aussi à la mise en lumière du caractère changeant, fluide, des données en amont de leur fixation (Lampland, 2010). Ce premier axe est évidemment central pour enquêter dans les coulisses de l'ouverture des données publiques. Comment le travail d'ouverture est-il organisé ? De quelles tâches se compose-t-il ?

Un autre aspect mis en avant par les travaux en STS concerne la nature même des résultats scientifiques et des données. Loin d'être universels par essence, ceux-ci sont toujours ancrés dans des écologies pratiques, orientés vers des problèmes particuliers. Ces éléments ne composent pas un contexte qui entoureraient les données en ne les touchant qu'à la marge : ils constituent un cadre de pertinence dont le détachement est toujours coûteux, voire risqué. Qu'ils soulèvent la question du savoir tacite (Collins, 2001) ou des communautés de pratiques (Lave & Wenger, 1991), les travaux sont nombreux à avoir adressé ce problème, selon des angles différents mais complémentaires. Ils montrent qu'une part des connaissances et des savoirs est dépendante de l'environnement dans lequel ils ont vu le jour. Les données, dans cette perspective, doivent être comprises comme *indexicales* au sens de la linguistique. Leur intelligibilité est intrinsèquement liée aux conditions locales de leur production et de leur usage.

Cette caractéristique a été mise en lumière avec une grande efficacité par les récents travaux qui ont étudié les nouvelles exigences de collaboration qui sont nées des grands projets scientifiques internationaux et interdisciplinaires. Ces projets se sont en effet confrontés à d'innombrables difficultés dès l'amont de la collaboration, au niveau même de l'échange des données. Outre les questions, organisationnelles, de contractualisation et de formes renouvelées de la propriété intellectuelle, ces résistances se sont jouées sur le plan des données elles-mêmes qui ne circulent pas naturellement d'un site de recherche à l'autre, d'une discipline à l'autre. À la vision de données transparentes qu'il suffirait de formater selon des standards adéquats pour assurer la collaboration, les STS ont opposé une vision indexicale, de données scientifiques dont la circulation et les échanges produisent ce que Edwards appelle des « frictions » (Edwards, 2010). Ce point reboucle sur le premier : si les données donnent lieu à des frictions, il faut reconnaître le travail supplémentaire que supposent leurs échanges et le

coût de la fabrique et de l'entretien des metadonnées dédiées à la fluidification de leur circulation (Baker & Bowker, 2007; P. Edwards *et al.*, 2011).

Enfin, les récents travaux rassemblés sous la bannière des « *infrastructure studies* » se sont également penchés sur la portée politique des données, en insistant sur les enjeux des classifications qui les fondent. Les grandes bases de données participent de la rigidification de l'information et ont des conséquences sur les phénomènes qu'elles recensent ou organisent. En stabilisant des formes de nominations, des catégories, etc. la collecte et la manipulation de données standardisées naturalisent des différences et, parce qu'elles nourrissent des infrastructures diffuses qui ne sont plus questionnées une fois mises en place, créent des exclusions quasi irréversibles, laissant de côté tout ce qui ne peut être réparti dans les classes distinctes qui président à l'organisation des données. Généralement tournées vers la simplification ontologique, les bases de données tendent à mettre en avant certaines entités et à en écarter d'autres (Bowker & Star, 1999). Cela est d'autant plus problématique qu'elles ont toujours un caractère performatif. Diffuses et opérationnelles, les bases de données font exister une forme de réalité « sous la main » de leurs utilisateurs, dont la pré-standardisation risque d'effacer une part de la multiplicité du réel (Bowker, 2000; Mol, 1999).

Les trois principales pistes d'analyse que dessinent les travaux présentés ici — l'épaisseur du travail de production des données, les frictions qu'implique leur circulation, et leur dimension politique — offrent un cadre utile pour interroger les conditions concrètes de l'ouverture des données publiques. Elles invitent à aller au-delà des assertions qui font des données un allant de soi, une ressource qu'il suffirait de libérer pour améliorer la transparence politique et stimuler l'innovation, afin d'explorer en détails les processus par lesquels passent les données avant d'être effectivement ouvertes, et avant même de devenir des données en tant que telles. En suivant cette invitation, nous proposons de nous focaliser ici sur trois aspects qui nous sont apparus essentiels dans le processus amont de l'ouverture : l'identification même des données à ouvrir, les problèmes que posent leur extraction et les transformations mises en œuvre pour assurer leur qualité de données brutes.

Identification et exploration

Dans les études qui portent sur les grands programmes d'échange de données scientifiques, l'existence même des données n'est jamais questionnée en tant que telle : ce sont les opérations sur les données et les métadonnées qui sont analysées et qui sont présentées comme les plus complexes, celles précisément qui mettent en lumière l'épaisseur des dites données et les frictions que provoque leur mise en circulation dans des espaces nouveaux.

Dans le cas des politiques d'open data, les données elles-mêmes ne vont pas de soi. Non seulement le périmètre des données potentiellement candidates à l'ouverture est débattu et travaillé pour chaque projet concret, mais la connaissance même de leur existence, de leur nature et de leur emplacement est également problématique. Avant la question « comment allons nous ouvrir tel ou tel jeu de données ? » de nombreux acteurs se confrontent à une interrogation plus abyssale encore : « de quelles données disposons-nous ? ». Dans le processus d'ouverture des données publiques, l'*identification* est ainsi une étape cruciale, complexe, qui, si elle se situe en amont du travail de production au sens strict, ne se résume que rarement au recensement plus ou moins fastidieux d'entités aux contours nets qu'il suffirait de débusquer dans l'organisation.

Dans la plupart des cas, les personnes chargées de la mise en œuvre d'un programme d'open data ne sont pas elles-mêmes productrices de données. Elles se trouvent donc, au démarrage du projet, en situation d'exploration, enquêtant par différents moyens sur l'existence potentielle de données candidates à l'ouverture. Cette exploration vise généralement à déboucher sur un inventaire, plus ou moins outillé, qui recense les données. Il existe une version utopique de cet inventaire, qui se présenterait sous la forme d'un catalogue exhaustif au sein duquel serait recensé, non pas les données dont l'ouverture serait jugée intéressante, mais « toutes » les données produites par l'institution.

> On aimerait bien dans l'idéal et partout d'ailleurs, c'est qu'il puisse avoir une liste absolue, exhaustive de toutes les données que produit chaque service de chaque collectivité publique. (…) S'il pouvait y avoir un annuaire complet, ça serait, c'est un peu utopique mais ça serait génial. Après, parmi toutes ces données, on identifierait celles qui peuvent être ouvertes, celles qui ne peuvent pas parce qu'il y a des restrictions, les données personnelles, les sensibles etc. (Chargé de projet open data dans une intercommunalité).

Cette vision d'un inventaire complet est proche de la métaphore pétrolière utilisée dans les injonctions à l'ouverture des données publiques lorsque celles-ci insistent sur une ressource présente mais inexploitée et mal connue au sein des administrations. Mais elle est ici assumée comme une utopie. Les équipes qui explorent les institutions ne se trouvent jamais confrontées à des données toutes prêtes qu'il suffirait de recenser. L'identification des données s'effectue progressivement, au fil d'échanges avec les services internes. Et le processus se nourrit lui-même, faisant émerger de nouvelles pistes au fur et à mesure de l'enquête qui fait découvrir aux personnes qui s'en chargent les méandres de l'institution.

> On descend, on descend jusqu'au plus petit dénominateur commun pour qu'on puisse identifier vraiment toutes les données. Et ce qui est fou, c'est qu'à partir de ces trente rendez-vous, à chaque fois que je les rencontre, ils m'identifient cinq autres personnes qu'il faudrait que je vois donc, en gros, c'est un peu exponentiel. (Chef de projet open data d'une collectivité locale).

Selon les situations que nous avons étudiées, la sollicitation de ces services prend des formes variées : appel à initiatives internes, contact direct, premières pistes de données pertinentes dessinées par l'équipe open data puis proposées aux services concernés, voire propositions faites par des associations de citoyens ou des développeurs qui ont eu l'occasion de les formuler.

> Quand on a commencé à leur parler de l'open data, les idées de données ont jailli à la fois du côté des services et puis, de notre côté, on avait déjà aussi un petit peu réfléchi à la question donc on avait déjà des choses à leur proposer : « Tiens vous pourriez ouvrir telle chose. » Par exemple, quand on est allé voir le service de prestations à la population, on savait déjà que ce qui serait facile pour eux à ouvrir et qui serait intéressant, c'était les données concernant les prénoms des enfants nés dans la commune. C'est une chose qui avait déjà été faite dans d'autres villes, donc on savait que ça pouvait être intéressant de les sortir aussi. On avait envie également de s'intéresser aux données concernant les élections parce que c'est quelque chose que la ville possède de manière assez précise, les chiffres par bureaux de vote etc., voilà. Et ça, il me semble que ça n'avait pas encore été trop fait donc justement on voulait et puis c'était la période, c'était juste avant les élections présidentielles et législatives donc, on s'est dit voilà, c'est un peu le moment de faire ça. Donc, c'est vrai, qu'on était arrivé vers eux en leur disant « voilà c'est vrai que ça, ça serait intéressant » mais c'était globalement

> un échange. Pareil, avec le service X, on est plutôt arrivé en leur disant « qu'est-ce qui est possible pour vous, d'ouvrir dans un premier temps etc. » On y est allé par étapes parce que tout n'était pas simple à faire. (Chargé de projet open data dans une intercommunalité).

On le voit, cet inventaire n'est en rien mécanique. Il repose sur des rencontres et des discussions, voire des négociations, qui lient étroitement des questions variées : l'alignement avec des pratiques existantes dans d'autres collectivités, l'opportunité au contraire de se distinguer à travers l'ouverture d'un jeu de données que personne n'a encore diffusé, l'intérêt que peuvent représenter certains données, les difficultés ou facilités techniques de leur diffusion, etc. Il faut ajouter à ces aspects l'un des éléments les plus saillants dans ses discussions, qui n'apparaît pas dans l'extrait d'entretien ci-dessus : la sensibilité. Au cœur des discussions entre les services internes et les équipes open data, la question du degré de sensibilité des données, c'est-à-dire du risque que leur ouverture représenterait est souvent cruciale.

> On [a identifié les données à ouvrir] avec le chef de projet [open data], qui est venu nous solliciter pour libérer nos données, et avec le directeur du génie urbain qui s'est engagé finalement à libérer les données produites par sa direction, avec une certaine qualité. Il y a eu une négociation pour voir les aboutissements de tout ça. (…) On a été rassuré par la possibilité de pouvoir négocier… C'est nous qui décidons de toute façon de la libération des données qu'on produit. C'est un peu notre propriété, même si ce sont des données du service public et donc, il pouvait y avoir des hésitations, notamment pour les problèmes de sécurité ou pour éviter que l'on tende un bâton pour se faire battre. Alors, j'ai deux exemples, en tête. Le câblage éclairage public. Dans la pratique, par exemple, on sait qu'il y a des types aux métiers pas très recommandables qui vont éteindre l'éclairage pour certains lampadaires en trifouillant dans les câbles et en enlevant les bons câbles histoire de faire du trafic tranquille, sans qu'il y ait trop de lumière pour les déranger. Alors si on libérait le câblage, ils pourraient savoir donc d'où vient le jus de telle armoire pour éclairer telle rue, péter telle armoire pour avoir tout un quartier dans le noir pour faire du trafic tranquillement. Bon, c'est un peu tiré par les cheveux mais voilà. Le risque existe et puis, bon, il n'y a pas vraiment d'intérêt aussi à libérer cette donnée. Alors, on a libéré uniquement le positionnement des mats.
>
> Une autre crainte c'était de libérer des données trop détaillées d'éclairage public. Notamment le modèle, date de pose, puissance des lampes, etc. ce qui pouvait donner des informations à des grands groupes privés qui gèrent les réseaux d'éclairage public, qui ont pour métier de gérer les réseaux d'éclairage public de certaines collectivités via notamment des partenariats publics-privés. (…) Il y avait une méfiance par rapport à ça. Donc, pareil, on a supprimé ces données-là de l'extraction qu'on en fait. (Gestionnaire de données dans le service d'urbanisme d'une municipalité)

Nous n'entrerons pas ici dans les détails de ces dimensions de sécurité et de sensibilité, qui mériteraient un article à part entière. Il est toutefois important de souligner la variété des enjeux que l'exploration et la mise en place de l'inventaire font apparaître. Si les données ne sont pas disponibles dans les services, toutes prêtes à être libérées, elles ne sont pas non plus sélectionnées sur la base de critères simples, définis à l'avance : leur exploration est progressive et collective. Elles passent par une co-construction qui montre que l'identification même des données qui vont être ouvertes participe pleinement du processus d'ouverture.

Deux points sont importants pour comprendre les enjeux de cette étape d'exploration. Tout d'abord, elle a des répercussions organisationnelles. La mise en place d'un inventaire n'est que rarement envisagée comme une opération unique, mais plutôt comme les premiers pas

d'un processus récurrent de circulation des données identifiées au sein de l'institution, des producteurs jusqu'aux dispositifs de leur ouverture. Un chargé de projet open data dans une collectivité locale explique ainsi qu'à terme les gestionnaires de données au sein des départements devraient arriver à maîtriser les enjeux de l'open data pour que la démarche perdure et ne nécessite pas l'intervention systématique de l'équipe en charge du projet :

> On est en train de réfléchir, parce que là on gère de la donnée quasiment au bout du tunnel. C'est-à-dire que la donnée, elle est connue quelque part, elle transvase dans différents services et nous, on la récupère, on l'extrait et on la met sur le portail avec des modifications. Et on est en train de se rendre compte qu'il faudrait pouvoir instaurer la culture de la donnée à la base. (Chef de projet open data dans une collectivité locale).

En stabilisant des circuits de diffusion, ce sont donc non seulement des types spécifiques de données qui sont identifiées, mais également des lieux dans l'institution et des personnes qui sont instituées en responsables de ces données et de leur circulation. Comme pour l'identification des données elles-mêmes, ce processus ne consiste pas en une description d'un organigramme déjà là, lisible, au sein duquel il suffirait de sélectionner des chemins de diffusion. La mise en place de l'open data « travaille l'organisation », pour reprendre les termes de Cochoy, Garrel et de Terssac (1998), redistribuant certaines cartes, attribuant des rôles nouveaux et des responsabilités inédites. Et plus l'inventaire devenu continu est équipé (plus il est inscrit dans des infrastructures techniques stables, jusqu'à devenir un outil de *workflow* dans certains cas), plus il ancre cette transformation organisationnelle dans le temps et en rigidifie les contours.

Deuxième aspect essentiel des processus d'exploration que nous avons brièvement décrits ici : ils montrent que l'identification des données est un geste d'instauration à part entière ; instauration aussi bien technique, qu'organisationnelle et politique. Les données ne sont pas révélées comme ouvrables, découvertes parmi une masse d'autres données déjà disponibles : elles sont instaurées en données à ouvrir, au fil de l'enquête menée en interne, de la confrontation des idées des uns et des autres, et des négociations dont nous venons de souligner la variété des arguments. Même lorsque l'exploration elle-même ne s'avère pas particulièrement complexe, cette instauration demeure essentielle dans le processus d'ouverture. Nous avons ainsi pu découvrir qu'au sein d'une organisation internationale seules certaines données très spécifiques avaient été identifiées dans l'inventaire réalisé pour l'open data. Ces données n'avaient pas été difficiles à désigner : elles émanaient du département qui était chargé du programme open data et faisaient déjà l'objet d'une publication dont une partie était jusqu'ici était payante. Même si elles apparaissaient comme évidentes aux yeux des personnes responsables du programme (ce qui constitue un cas très particulier dans notre étude), ce choix opérait malgré tout lui aussi une instauration de ces données — et ces données seulement — comme données ouvertes. Ainsi, dans les documents qui ont circulé lors de la mise en place du projet, les informations qui ne relevaient pas des données déjà publiées n'étaient pas écartées du processus d'ouverture à l'issue de négociations particulières ou de choix politiques argumentées, mais simplement qualifiées de *non-data* sans autre forme de procès. Qu'elle résulte d'une exploration au long cours, ou d'une désignation fluide et quasi « naturelle », l'opération d'identification est donc générative. Elle engendre une certaine réalité (Law, 2009), un périmètre de données qui sont instaurées non seulement comme « ouvertes » (ouvrables, dans un premier temps), mais aussi comme « données » tout court.

Extraction

Une fois les données candidates à l'ouverture identifiées, il faut être capable de mettre la main dessus, ce qui ne va pas non plus de soi. Les données ne sont pas devenues disponibles du simple fait de leur recensement : elles restent encapsulées dans les bases de données et leur mise à disposition nécessite qu'elles en soient extraites. Pour la plupart de leurs utilisateurs, l'accessibilité des données est en effet toujours subordonnée aux logiciels qui les rendent visibles, les ordonnent. Ces logiciels produisent des interfaces métiers, ce qui s'appelle dans le vocabulaire des bases de données relationnelles des « vues utilisateurs », destinées à simplifier l'usage de la base et dont la multiplication permet de faciliter la variété même des usages (Dagiral & Peerbaye, 2013). Ces vues ne donnent pas à voir l'organisation physique des données dont elles « protègent » les utilisateurs (Castelle, 2013). Les usagers sont donc dépendants de ces vues pour entrer en contact avec les données et rares sont les bases de données qui sont affublées d'une fonctionnalité d'extraction automatique, qui permettrait de récolter les données indépendamment de leur mise en forme logicielle.

> Il n'y a aucun [logiciel] qui intègre ça dès le départ dans le produit avec effectivement les informations à extraire. Souvent, [les prestataires] sont effectivement propriétaires de leurs schémas. C'est-à- dire qu'en fait, ce sont des gens chez eux qui ont développés donc qui ont réfléchis comment ils organisent les données, comment ils les présentent etc. et donc, du coup, même s'ils travaillent pour des collectivités, la plupart ne vendent pas qu'à des collectivités. Donc, ils ne sont pas concernés par l'open data. (Database Manager dans une municipalité)

Après l'exploration dédiée au recensement des données en interne, c'est donc une autre exploration qui démarre, à travers laquelle les informaticiens cherchent à récupérer les données elles-mêmes, leurs modalités de stockage et leur organisation. Pour assurer cette extraction, il faut fouiller au-delà des interfaces de visualisation, dans les entrailles des serveurs, à la racine même de la base de données. Il faut se défaire des « vues utilisateurs » pour reconstituer la « vue physique » de la base de donnée, et développer un outil (une « moulinette ») qui, à partir de cette reconstitution, pourra accomplir l'extraction.

> En fait, ce qui est complexe, il faut bien comprendre c'est qu'au départ pour la plupart des systèmes des applications qu'on a chez nous qu'on a acheté, elles ne sont pas du tout conçues pour faire de l'open data. Donc, c'est compliqué. On est obligé, nous, de développer des moulinettes, des tas de choses pour pouvoir sortir des données proprement. (Gestionnaire de bases de données de transport).

> C'est une requête SQL (…) qui s'attaque à la base de données de surf parce que SPORT [le système qui fournit les vues utilisateurs], qui étrangement n'a pas été conçu pour sortir directement ces tableaux, ces tableaux de données agrégées. Il est conçu pour sortir des tableaux, le tableau mensuel d'un capteur mais pas tous les capteurs en table dans un dans un seul tableau Excel. Bon, il a été conçu comme ça mais en même temps, c'est pas grave parce qu'on sait le faire malgré tout via une requête SQL (…). C'est moi qui ai développé cette demande, et les personnes qui s'occupent du système SPORT m'ont fait des requêtes spécifiques pour aller attaquer la base et agréger des données telles que je les voulais. (Gestionnaire de données de voirie).

Or, même si de grands principes se retrouvent dans ces modalités de rangement, dans la manière dont les données sont effectivement stockées sur les disques durs, cette « vue

physique » est toujours spécifique, comme le sont les manières d'organiser les placards personnels.

> Ce qu'il faut te dire c'est que rien n'est universel là-dedans. C'est-à-dire que la manière dont tu ranges tes données c'est comme la manière dont tu ranges tes chaussettes à la maison, chacun peut les ranger de manière différente. On a tous le même placard, mais on les range tous de manières différentes. (Database Manager dans une municipalité)

Chaque outil d'extraction est donc toujours fait sur mesure, et le travail est d'autant plus complexe que les bases de données et les logiciels qui y donnent accès, voire les différentes versions d'un même logiciel, se sont accumulés au sein des institutions. C'est parfois une véritable foule d'instruments à laquelle les informaticiens ont affaire, dont l'exploration représente un coût très important.

> Ce qui peut aussi poser problème (…) c'est que chaque logiciel étant unique, les formats de données sont tous différents, et les schémas de répartitions des données sont tous différents, donc une procédure que tu as utilisée pour un logiciel ça sera pas la même pour un autre, même si tu reprends un peu les bases. Le corps est à peu près le même mais les informations, elles, ne seront pas stockées de la même manière, donc il faudra refaire ce processus d'analyse pour chaque base de données différente. Et des bases de données, on doit en avoir peut-être au moins cinquante différentes. Donc, c'est extrêmement long d'extraire ces données là. À la mairie, on a des données depuis plus de trente ans, qui en plus sont arrivées à l'époque sur les grands systèmes IBM qui sont différents des systèmes Windows, qui sont différents des systèmes Linux. On n'a à peu près de tout à la mairie. Du coup, c'est très compliqué d'extraire quelque chose de précis. (Database Manager dans une municipalité)

Ce travail d'extraction constitue l'opérationalisation de l'identification des données, deuxième étape de leur progressive instauration. Il montre un autre aspect des données brutes, dont, on le voit, la disponibilité n'est pas une propriété intrinsèque. Pour mettre en œuvre une politique d'open data, il faut être capable, une fois les données identifiées, de contourner les bases de données d'une manière ou d'une autre et récolter les données à même leur espace de stockage. Les bricolages mis en œuvre pour cela, les explorations et les « moulinettes » qui permettent d'atteindre les données, donnent une idée de l'épaisseur de la nasse sociotechnique dont celles-ci doivent être littéralement extirpées. D'autant que la lutte avec les logiciels qui fournissent les vues utilisateurs sur les bases de données ne soulèvent jamais des questions « purement » informatiques. Les explorations mises en œuvre pour l'extraction des données mettent directement en jeu les relations entre les départements informatiques concernés et leurs prestataires de service.

> C'est un logiciel produit par une boîte américaine qui a quelque chose comme trois clients en France et qui ne s'en occupe pas beaucoup. Et c'est un logiciel qui est complètement opaque. C'est-à-dire que nos équipes ne maîtrisent pas du tout ce qu'il y a dedans, ce que fait le logiciel et ce qu'il peut sortir à la fin. Et elles ne peuvent pas trop y toucher. C'est-à-dire qu'elles n'ont pas, par exemple, d'accès direct à la base de données. Elles sont obligées de passer par le formulaire que leur a gentiment fourni le prestataire. On peut considérer que les données leur appartiennent et pourtant, à cause de ce logiciel qui les bride un petit peu et bien elles n'ont pas pu faire tout ce qu'elles voulaient. Par exemple, au service des jardins, ils avaient envie de mettre en place un système pour pouvoir accéder directement, en temps réel, au contenu des pépinières, les plantes qui y sont, toutes les

> informations sur les plantes etc. mais également l'avancement de leur culture. Est-ce qu'elles sont prêtes à être plantées ou est-ce qu'elles sont encore en culture ? et on s'est rendu compte que c'était compliqué parce que comme on n'a pas l'accès à la base de données et bien on peut pas, on peut pas aller piocher dedans comme on veut. Donc, justement, on est en train de travailler avec eux pour arriver à détourner ce système, pour pouvoir quand même aller interroger la base de données mais ça n'est pas évident. (Chargé de projet open data dans une intercommunalité).

Selon les termes des contrats, et la plus ou moins bonne volonté des prestataires, les chemins d'accès vers les données « elles-mêmes » sont plus ou moins difficiles à obtenir, et les bricolages pour y mener plus ou moins assimilables à des détournements, voire à des ruptures contractuelles. Ce point est essentiel pour remettre en question l'idée selon laquelle les données publiques seraient des ressources dormantes qui ne demanderaient qu'à être libérées pour être exploitées. Cette vision de la donnée comme *« commodity »* (Ribes & Jackson, Steven, 2013) est mise à mal non seulement par le coût que représente le travail d'extraction, mais aussi par l'ambiguïté de ce que les prestataires techniques mettent à la disposition des institutions en leur fournissant leur système de bases de données. Certains prestataires sont propriétaires des chemins d'accès et des systèmes de stockage de leurs bases de données : l'inaccessibilité des données est au cœur de leur modèle économique.

> Mais comme il y a beaucoup, beaucoup de logiciels ou de progiciel « propriétaires » (…), les schémas de bases de données appartiennent aux sociétés, les configurations appartiennent aux sociétés. Même si on met ça sur notre matériel, on n'est pas libre de faire ce qu'on veut. (Database Manager dans une municipalité)

D'un certain point de vue, le travail d'extraction consiste donc pour les institutions à reprendre la main sur les données, en désarticulant de manière très concrète les assemblages sociotechniques qui les lient à certaines entreprises privées. Plus généralement, cette désarticulation donne à voir, sous un angle très pratique, le feuilletage des infrastructures informationnelles. Comme l'ont montré Star et Ruhleder (1996), toute infrastructure repose sur une autre et est prise dans des jeux d'interdépendance complexes qui rendent délicate toute opération qui viserait à la singulariser. L'ouverture des données publiques passe par ce type de singularisation, qui est bien entendu momentané. Les données sont extraites, isolées des bases qui les ordonnaient et assuraient leur accessibilité ordinaire, afin d'être déplacées et inscrites dans un nouvel assemblage, dédié à leur ouverture.

Mais le travail d'ouverture de ne s'arrête pas aux opérations d'identification et d'extraction. Si ces deux étapes, étroitement liées, sont cruciales dans l'instauration progressive des données ouvertes, elles ne suffisent pas à leur transformation. Pour être ouvertes, les données doivent être elles-mêmes travaillés, façonnées.

« Brutification »

« Données brutes » est un terme ambigu qui, lorsque les pratiques qui y sont associées sont étudiées de près, s'apparente à un oxymore (Bowker, 2000; Gitelman, 2013). C'est flagrant à propos des projets open data dont l'analyse donne à voir la série de transformations, de raffinements pour reprendre à notre compte la métaphore pétrolière, dont les données sont l'objet pour *devenir* brutes. Trois grands types d'opérations peuvent être distinguée pour décrire les transformations : le reformatage des données, leur nettoyage et leur désindexicalisation.

Nous l'avons évoqué à propos des démarches d'extraction, les données existent au sein des institutions sous des formes variées, accessibles dans des formats techniques hétérogènes. Cependant, leur extraction ne suffit que rarement à en faire des données ouvertes. Leur ouverture doit passer par un formatage qui assure qu'elles soient lisibles par les outils les plus communs des développeurs, mais aussi du grand public. Évidemment, cette seule idée d'un format « convenable » cache une complexité que nous ne déplierons pas ici. Les bons formats pour l'open data sont abondamment discutés et la standardisation des modes de diffusion des données ouvertes sur ce plan en est encore à ses balbutiements. Ce qu'il est important de retenir ici est que cette question des formats occupe un pan important des transformations que les données subissent pour être effectivement considérées comme ouvertes. Au cours de notre enquête, le format CSV (comma-separated values) était le plus fréquent, choisi notamment parce qu'en tant que format ouvert, il permet une utilisation par les logiciels tableurs les plus répandus. On a ainsi pu entendre dans une réunion de démarrage du programme open data d'une grande entreprise publique cette phrase qui souligne l'importance de la mise au format dans le processus d'instauration des données ouvertes : « pour moi du brut c'est du csv ».

Or, la traduction d'un jeu de données en fichier CSV ne va pas de soi. Chaque exportation d'un format propriétaire issu des bases de données originales, ou d'un logiciel tableur, vers le format désiré réserve son lot de problèmes et donne lieu à des ajustements qui assurent, autant que faire se peut, que les données initiales ne soient pas corrompues par le processus. Dans le meilleur des cas, des opérations en amont de l'exportation permettent de produire des fichiers qui seront compatibles et supporteront mieux le reformatage.

Avant ou après leur mise au format, les données subissent un autre type de traitement : elles sont « nettoyés ». Le vocabulaire du nettoyage, associé à l'idée de « mise en qualité » des données est largement employé dans les domaines scientifiques. Il a récemment fait l'objet d'une thèse qui en explore les richesses à partir d'une ethnographie de travaux scientifiques menés en forêt amazonienne (Walford, 2013). Dans le cas de l'open data, le nettoyage concerne plusieurs aspects. Il consiste d'abord à corriger les erreurs au sein des jeux de données : les valeurs qui sont repérées comme aberrantes, mais aussi les « trous » dans les fichiers (l'absence de valeur). Le nettoyage implique également l'harmonisation des données. Comme nous l'avons vu, les jeux de données se côtoient au sein d'une même institution, dans des formats différents, manipulés par des services qui les produisent et traitent de manière spécifique. Ainsi, des entités *a priori* identiques peuvent-elles avoir dans les bases de données de l'administration des unités de valeurs, voire des identifiants, différents. Comme dans le cas des partages de données scientifiques à grande échelle (Millerand, 2012) la production de jeux de données ouvertes passe par la mise en cohérence de ces écarts et éventuelles redondances.

> Typiquement, sur les jeux de données des élections : entre les derniers fichiers des dernières élections, et puis les vieux trucs (on est remonté jusqu'à 2004), les fichiers n'étaient pas présentes pareil. C'était des choses très bêtes mais il y avait des fois le titre de colonne qui était soit le nom du candidat soit le nom de son parti ou alors les deux, et j'ai essayé d'uniformiser tout ca pour que tous les fichiers se ressemblent et soient structurés pareil. (Chargé de projet open data dans une intercommunalité).

C'est un aspect important de l'open data. Les programmes d'ouverture mettent à l'épreuve des données qui, si elles étaient publiées telles quelles, pourraient passer pour des données de mauvaise qualité, alors même que leurs usagers en interne n'y voyaient rien à redire jusque là.

On retrouve ici une question largement discutées en STS et au-delà, à propos des « bad records » (Garfinkel, 1967) et des « false numbers » (Lampland, 2010). Au sein des organisations, les données dites « métiers » ne sont pas justes ou vraies en elles-mêmes. Leur faible degré de précision, ou ici leur manque d'harmonisation, n'ont aucun impact sur leur efficacité, au contraire. Il existe de nombreuses bonnes raisons organisationnelles, pour reprendre les termes de Garfinkel, pour que ces données persistent, tout simplement parce que leur justesse et même leur « vérité » sont ancrées dans les pratiques de ceux qui les manipulent et les mobilisent. C'est la confrontation entre des domaines de pratiques aux enjeux différents qui peut amener à stigmatiser telle ou telle donnée comme fausse ou mauvaise. Les programmes d'ouverture des données publiques constituent de ce point de vue des épreuves. En faisant migrer les données dans un cadre nouveau, ils rendent potentiellement centrales certaines de leurs dimensions qui étaient peu pertinentes dans leurs cadres d'usages initiaux. Des absences jamais remarquées deviennent des manquements, des approximations ou des doublons sans importance deviennent des erreurs ou des redondances. Dans tous les cas, le passage d'un cadre à un autre s'accomplit par ces opérations de nettoyage.

L'idée même de nettoyage met l'accent sur un aspect essentiel du cadre vers lequel les données sont destinées à migrer. Nettoyées, les données deviennent génériques, débarrassées de scories issues des activités situées qui s'en nourrissaient. Le nettoyage est un premier pas vers l'universalité que l'open data, comme beaucoup d'autres projets[4], suppose. L'effacement des ambiguïtés et la disparition des absences produisent des jeux de données propres à tous les usages virtuellement possibles. Cette projection des données vers l'universalité des usages possibles est alimentée par une autre étape, très proche du nettoyage : la désindexicalisation. Outre la production de données auxquelles rien ne pourra être reproché en termes de cohérence et de complétude, le travail de « brutification » passe également par l'effacement des traces d'usages précédents.

> Typiquement, pour les statistiques de fréquentation, par exemple, c'était le fichier de travail [du département]. C'était un fichier Excel qu'ils avaient mis en forme selon ce dont ils avaient besoin. Ils avaient fait un tableau avec leur propre titre de colonnes, des couleurs… On sentait vraiment que c'était fait parce que ce fichier Excel ils travaillaient dessus. Il faut dire que leur processus est un peu compliqué. En gros, le logiciel leur pond des chiffres et ils les prennent à la main pour le mettre dans un fichier pour établir les statistiques globales. Donc, c'était vraiment leur fichier de travail. Or, nous, on ne voulait pas ça. Nous, on voulait des données plus brutes c'est-à-dire pas de commentaires, pas de tableaux, pas de mise en forme, juste vraiment les données au jour le jour, statistiques. Moi, je me suis occupée de ce travail là, rebrutifier les données en fait, pour qu'elles soient vraiment le plus simple possible à utiliser ensuite pour les développeurs. (Chargé de projet open data dans une intercommunalité).

Pour devenir ouvertes, les données doivent être délocalisées, ne plus porter en elles les marques de leur ancrage professionnel. Cette étape constitue une autre dimension de l'instauration des données, dont le périmètre est à nouveau spécifié par la mise à l'écart d'éléments spécifiques (les couleurs, les commentaires, mais aussi parfois le découpage en

[4] Les contributeurs d'OpenStreetMap, outil de cartographie collaborative, travaillent par exemple à produire des données géographiques les plus universelles possibles en se débarrassant autant que faire se peut de toutes les catégories qui répondraient à des pratiques ou des usagers spécifiques 14/05/2014 10:07.

sous-catégories, etc.) qui de fait sont considérés comme ne faisant pas partie des données « en tant que telles ». Dans la suite du nettoyage, la désindexicalisation participe donc à une purification des données. Mais elle met également en œuvre une autre dimension centrale aux programmes d'open data et à la politique de transparence qu'ils servent : leur mise en intelligibilité. Une part des aspects « métiers » des jeux de données candidats à l'ouverture pose en effet des problèmes de compréhension qui vont bien au-delà du parasitage que pourrait représenter aux yeux des personnes en charge de la campagne open data des couleurs ou des mises en formes locales. Le cas le plus flagrant et le plus fréquent porte sur les acronymes et les abréviations. Les écrits professionnels sont pour une grande part des écrits abrégés (Fraenkel, 1994) : toute organisation repose sur ces formes langagières réduites, plus ou moins foisonnantes, souvent moquées par les non initiés, par lesquelles de nombreuses entités sont identifiées de manière opératoire. Ces acronymes et abréviations sont traités comme des brèches dans le processus d'ouverture des données, qu'il faut réparer.

> À quatre-vingt-dix pour cent ce sont des données purement techniques avec par exemple, les libellés commerciaux au lieu que ce soit marqué « Boulevard du Général de Gaule » c'est marqué « Bd DGL » par exemple. Parce que c'est un code qui suffit largement quand les départements concernés conçoivent les horaires, « Bd DGL », ils savent à quoi ça correspond. Le voyageur ça ne lui parle pas, ça ne lui parle pas du tout, donc il a fallu pouvoir croiser certaines bases chez nous qui ont les bons libellés. Aujourd'hui, pour construire le lot GTFS [General Transit Feed Specification, norme de données de transport] on croise avec six à sept bases. (Gestionnaire de bases de données de transport).

En parallèle d'une dimension tournée vers la propreté et la pureté des données, le processus de « brutification » s'attèle ainsi à leur intelligibilité. S'il faut travailler les données pour en faire des données brutes, c'est donc bien, encore une fois, qu'elles ne sont pas disponibles en soi : elles ne portent pas en elles une intelligibilité nécessaire au cadre de l'open data. Cette mise en intelligibilité passe évidemment par l'élaboration de métadonnées essentielles à tout projet de partage de données (Bowker et Baker, 2007; Edwards *et al.*, 2011) : dictionnaire, commentaires dans des documents à part, sont associées aux données brutes pour que leurs usages soient facilités. Mais elle passe aussi par la transformation des jeux de données eux-mêmes, au sein duquel des termes vont être remplacés, des intitulés simplifiés, d'autres assemblés.

Les données brutes qui sont jugées bonnes pour l'open data sont donc des données au bon format, « propres » et intelligibles. À l'instar de l'identification et de l'extraction, la « brutification » qui est mise en œuvre pour atteindre ces qualités a des ancrages organisationnels importants. Comme dans le cas de projets de collaboration scientifique, elle suppose d'abord que certaines personnes prennent en charge un certains nombre de tâches qui n'ont généralement pas été pensées et qui nécessitent que s'inventent au cas par cas des compétences et des places dans la division du travail (Millerand 2012). Mais elle peut aussi avoir des conséquences sur l'organisation elle-même. Pour certaines institutions, les étapes que nous venons de décrire sont en effet l'occasion de repenser des process organisationnels, afin de minimiser le travail en aval sur les données. Cela passe par de nouvelles formes de collaboration, ou par la mise en place de nouvelles étapes dans la gestion initiale des données.

> Sur les arrêts [de bus], on en a profité pour travailler avec la municipalité qui a aussi sa propre base d'arrêts et on en a profité avec eux pour uniformiser nos bases et avoir les mêmes données dedans. On a travaillé avec eux pour que tous les noms d'arrêts soient identiques chez eux comme chez nous.

> Donc, maintenant on les a mis dans la chaîne, comme ça toutes les bases sont à jour en même temps. Les gens ne voient pas que l'effet de bord de l'open data c'est que ça a permis de fiabiliser notre système d'information et notre qualité de données. C'est primordial pour pouvoir développer des nouveaux systèmes d'information. [L'open data] a plein d'impacts qu'on n'imaginait pas forcément au départ. (…)
>
> Des fois, c'est juste qu'il manquait de communication avec d'autres services. On s'est rendu compte que, par exemple, pour le changement d'un nom d'arrêt, il n'y avait pas forcément de communication de la part de la personne qui faisait changer le nom d'arrêt. Elle ne redescendait pas toujours l'info au service qui concevait les horaires et c'est pour ça que des fois, le nom d'arrêt dans mon fichier n'était pas bon parce qu'on ne m'avait pas communiqué l'info. Alors, j''ai identifié où étaient les problèmes et après j'ai fait remonter aux différentes personnes qui, entre elles, ont remis en place un process d'information « quand je change un nom d'arrêt, j'envoie un mail à un tel ». Voilà c'est tout bête mais comme avant ça ne se voyait pas, ce n'était pas grave, le nom d'arrêt à la limite on s'en moque du moment qu'on arrive à concevoir les horaires. Il y a donc eu un travail qui a été fait de cartographier un peu le processus et mettre en évidence ce qui peut-être n'allait pas pour l'améliorer. (Gestionnaire de bases de données de transport).

Cette intégration d'une partie du travail de « brutification » dans l'organisation du travail des instituions est au cœur des projets les plus récents qui voient dans la politique d'open data, outre son ancrage dans une politique de la transparence, un levier de modernisation de l'administration. Il faut toutefois bien en prendre la mesure. Ces réorganisations ne relèvent pas seulement de la reconnaissance d'un travail sur les données, jusque là impensé, que l'on inscrit désormais dans les process internes. Elles consistent également à renverser le mouvement de « brutification » tel que nous l'avons décrit, en intégrant les problématiques d'ouverture dans les activités qui jusque-là bénéficiaient de données *ad hoc*. Uniformiser, nettoyer et désindexicaliser en amont, revient à faire travailler chacun avec des données déjà brutes, c'est-à-dire des données génériques qui perdent les qualités de leur ancrage. Autrement dit, il s'agit de faire de la transparence non plus le résultat d'opérations spécifiques, mais l'horizon même des activités administratives, indépendamment des particularités des métiers et des données de chacun. Ce renversement se fait au risque de retrouver les situations décrites par Garfinkel (1987) de décalage entre registres de pertinence peu compatibles : celui de la recherche *versus* celui de la gestion des soins dans le cas de Garfinkel, celui de l'ouverture générique *versus* celui de l'ancrage pratique dans notre cas.

Les situations de réorganisation interne, que nous ne faisons qu'évoquer ici, montrent qu'il existe deux grandes directions possibles pour prendre en considération le travail de fabrication des données brutes. Une fois ce travail éprouvé, puis reconnu, c'est-à-dire une fois que l'on assume que l'ouverture des données à un coût, qu'elle représente même un investissement, on peut l'assumer comme une série d'opérations à mener *a posteriori* sur les données métier, il faut alors inventer des postes et redéfinir des rôles au sein de l'organisation. On peut au contraire chercher à intégrer ce travail en amont, en transformant la nature même des données sur les sites de leur production et dans leurs premiers usages. La différence entre les deux directions ne tient pas tant à la part organisationnelle de la fabrique des données brutes (elle est présente à chaque fois), mais à la définition sous-jacente de ce que l'on entend par données. Dans le premier cas, la multiplicité des données et la nécessité d'en faire coexister des versions différentes au sein de l'institution sont assumées. Dans le second, le caractère générique des données — leur aspect « brut » — est considéré comme un bien en soi, sur lequel il faut aligner les idiosyncrasies professionnelles.

Conclusion

Au cours de cet article, nous avons cherché à montrer qu'au contraire de ce que certaines injonctions, voire certaines dispositions légales, pouvaient le laisser entendre, les données des administrations publiques ne sont pas des entités « déjà là », prédisposées à une ouverture qui tiendrait dans une simple mécanique de libération. L'ethnographie des activités concrètes sur lesquelles reposent les projets open data montre que les données sont progressivement instaurées en données ouvertes. Nous avons insisté sur trois étapes cruciales de cette instauration — l'identification, l'extraction, et la « brutification », qui constituent une part du travail sociotechnique dont les données font l'objet. En sciences, ces transformations ont été étudiées comme les opérateurs de mutation des données brutes en données certifiées, prêtes à être mobilisées dans des activités spécialisées (Edwards, 1999; Walford, 2013). Dans notre cas, nous l'avons vu, l'enjeu est de passer de données métier à des données brutes. D'un certain point de vue, le mouvement est inversé et par là même le sens des données brutes diffère. Comme l'explique A. Walford en détails, les données brutes en sciences sont des données en devenir, des entités ambiguës, multiples, qui « attendent » la série des traitements qui vont assurer la consolidation progressive de résultats scientifiques par leur inscription dans un réseau sociotechnique stabilisé (Walford 2013). Le travail qui est effectué sur elles en amont de ces traitements consiste avant tout à les aligner à quelques pré-requis largement partagés par des opérations de nettoyage semblables à celles que nous avons évoquées plus haut. Dans les programmes open data, les données brutes ne peuvent être un point de départ. Elles ont déjà « vécu », elles sont déjà inscrites dans des réseaux sociotechniques qui les stabilisent et les orientent vers des pratiques spécifiques. Les données brutes de l'open data sont le résultat d'opérations qui visent à désencastrer les données de ces réseaux, à les débarrasser de la gangue des pratiques qui en faisaient des données métier, pour les transformer en données ambiguës, « ouvertes » à de nombreux types de traitements. Autrement dit, les données sont brutes lorsque l'on réussit à les dé-spécifier de leurs usages initiaux pour les préparer à un vaste horizon d'usages possibles. L'enjeu des transformations qu'elles subissent n'est plus de corriger des biais de mesure ou de distinguer le bruit de l'information à traiter (comme en sciences), mais d'assurer une migration d'une donnée « étroite », locale, vers une donnée à vocation universelle. Il s'agit en quelque sorte d'assurer le passage entre les deux types de données discutées par A. Desrosières : les sources administratives, toujours territorialisées, et les données statistiques à vocation de généralisation (Desrosières, 2005). Bien entendu, ce désencastrement des données métier ne suppose pas que les données puissent exister par elles-mêmes, une fois « libérées ». Il passe au contraire par un nouvel encastrement, celui de l'uniformisation, des formats standards, etc. : l'inscription dans un nouveau réseau sociotechnique qui opère ses propres formes de réduction et de clôture.

Pour conclure, il nous faut maintenant revenir sur le vocabulaire même de la donnée brute. Doit-on, une fois mises en lumière les opérations qui participent à leur façonnage, considérer qu'il n'existe donc « en réalité » pas de données brutes, et que la notion même relève de la fiction, voire de l'illusion ? Sans doute pas, à moins que l'on cherche à imposer une définition de la donnée brute qui n'est finalement pas partagée par les acteurs concernés. Si l'on peut identifier une revendication de la donnée déjà-là, commodité à portée de main qu'il suffit de libérer, dans les injonctions à la libération et les promesses qui les nourrissent (Goëta, 2012), cette définition n'est pas revendiquée en tant que telle en interne par les producteurs et les manipulateurs de données, qui font l'expérience au jour le jour du travail nécessaire à la transformation des données métier en données brutes, et l'assument. C'est évidemment le cas des pratiques de ce que nous avons appelé la « brutification », mais également des opérations d'identification et d'extraction. Plutôt que le vocabulaire de la fiction, il nous semble donc

préférable d'adopter celui de l'oxymore (Bowker 2000, Gitelman 2013), qui permet de mieux rendre compte des tensions que représente pour les travailleurs de l'open data cette transformation des données dont le coût n'est jamais complètement mesuré à l'avance.

Une telle posture pragmatiste, qui ne dessaisit pas les acteurs de leur vocabulaire et explore avec eux les méandres de la fabrique progressive des données brutes, invite à revenir sur le terme même de « donnée » et ses sens multiples, notamment en français. Rosenberg a récemment montré les liens entre la donnée au sens mathématique et ce qui « donné » au sens à la fois de déjà là et de postulat rhétorique. La donnée en sciences a pendant longtemps désigné le matériau de départ de l'analyse, du calcul, sans que son rapport à la réalité ne soit pris en considération : la donnée était un point de départ du travail scientifique, ce avec quoi il fallait faire, et pas forcément un bon représentant du réel (Rosenberg, 2013). Étudier les pratiques de l'open data de l'intérieur, comme nous l'avons fait, invite à faire un autre rapprochement. Si les données ne sont pas des points de départ pour les responsables des programmes que nous avons accompagnés, elles sont en revanche mises à la disposition d'autres utilisateurs. Les données, en ce sens, s'apparentent à des *dons*, des matériaux de départ offerts à la collectivité. Dans une certaine mesure, nous avons cherché à montrer que cette mise à disposition ne relevait pas de la livraison de données déjà existantes qu'il s'agirait de restituer, mais d'un travail méticuleux de façonnage de données que l'on s'engage à donner.

Références

Baker, K. S., & Bowker, G. C. (2007). Information ecology: open system environment for data, memories, and knowing. Journal of Intelligent Information Systems, 29(1), 127–144.

Baker, K. S., & Millerand, F. (2009). Infrastructuring ecology: challenges in achieving data sharing. In E. J. Parker, N. Vermeulen, & B. Penders (Eds.), Collaboration in the New Life Sciences (pp. 1–37).

Barley, S. R., & Bechky, B. A. (1994). In the Backrooms of Science. Work & Occupations, 21, 85–126.

Barry, A. (2001). Political Machines. Governing a technological society. New York: The Athlone Press.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. Science, (323), 1297–1298.

Bowker, G. C. (2000). Biodiversity Datadiversity. Social Studies of Science, 30(5), 643–683.

Bowker, G. C., & Star, S. L. (1999). Sorting Things Out: Classification and Its Consequences. MIT Press.

Castelle, M. (2013). Relational and Non-Relational Models in the Entextualization of Bureaucracy. Computational Culture, (3).

Collins, H. M. (2001). Tacit Knowledge, Trust and the Q of Sapphire. Social Studies of Science, 31(1), 71–85.

Dagiral, É., & Peerbaye, A. (2013). Voir pour savoir. Concevoir et partager des « vues » à travers une base de données médicales. Réseaux, (178-179), 163–196.

Denis, J., & Pontille, D. (2012). Les travailleurs invisibles de l'information. Revue d'anthropologie des connaissances, 6(1).

Denis, J., & Pontille, D. (2013). Une infrastructure élusive. Aménagements cyclables et troubles de la description dans OpenStreetMap. Réseaux, (178).

Desrosières, A. (2005). Décrire l'État ou explorer la société : les deux sources de la statistique publique. Genèses, (58), 4–27.

Edwards, P., Mayernik, M. S., Batcheller, A., Bowker, G., & Borgman, C. (2011). Science Friction: Data, Metadata, and Collaboration. Social Studies of Science, 41(5), 667–690.

Edwards, P. N. (1999). Global climate science, uncertainty and politics: Data-laden models, model-filtered data. Science as Culture, 8(4), 437–472.

Edwards, P. N. (2010). A Vast Machine. Computer Models, Climate Data, and the Politics of Global Warming. Cambridge: MIT Press.

Edwards, P. N., Bowker, G. C., Jackson, Steven, J., & Williams, R. (2009). Introduction : An Agenda for Infrastructure Studies. Journal of the Association for Information Systems Introduction, 10, 364–374.

Fraenkel, B. (1994). Le style abrégé des écrits de travail. Cahiers du français contemporain, (1), 177–194.

Garfinkel, H. (1967). Studies in ethnomethodology. Englewood-cliffs: Prentice-Hall.

Gitelman, L. (Ed.). (2013). "Raw Data" is an Oxymoron. Cambridge: MIT Press.

Goëta, S. (2012). Open data. Qu'ouvre-t-on avec les données publiques ? Mémoire M2, CELSA.

Goffman, E. (1973). La mise en scène de la vie quotidienne 1 : La présentation de soi (Minuit.). Paris.

Hine, C. (2006). Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work. Social Studies of Science, 36(2), 269–298.

Knorr-Cetina, K. (1981). The manufacture of knowledge. An essay on the constructivist and contextual nature of science. Oxford: Pergamon Press.

Lampland, M. (2010). False numbers as formalizing practices. Social Studies of Science, 40(3), 377–404.

Latour, B. (1985). Les « Vues » de l'esprit. Une introduction à l'anthropologie des sciences et des techniques. Culture Technique, 14, 4–29.

Latour, B., & Woolgar, S. (1988). La vie de laboratoire. Paris: La Découverte.

Lave, J., & Wenger, E. (1991). Situated Learning: Legitimate Peripheral Participation. Cambridge University Press.

Law, J. (2009). Seeing Like a Survey. Cultural Sociology, 3(2), 239–256.

Lynch, M. (1985). Art and Artifact in Laboratory Science. A Study of Shop Work and Shop Talk in a Research Laboratory. London: Routledge, communication series.

Lynch, M. (1993). Scientific Practice and Ordinary Action. Cambridge: Cambridge University Press.

Millerand, F. (2012). La science en réseau. Les gestionnaires d'information « invisibles » dans la production d'une base de données scientifiques. Revue d'anthropologie des connaissances, 6(1), 163–190.

Mol, A. (1999). Ontological politics. A word and some questions. In John Law & J. Hassard (Eds.), Actor Network Theory and After (pp. 74–89). Oxford: Wiley-Blackwell.

Pontille, D. (2013). Le vocabulaire de la contribution. Formes d'attribution et écologies du travail scientifique. Mémoire pour l'Habilitation à diriger les recherches, EHESS.

Power, M. (1997). The Audit Society: Rituals of Verification (p. 200). Oxford: Oxford University Press.

Ribes, D., & Jackson, Steven, J. (2013). Data Bite Man: The Work of Sustaining a Long-Term Study. In L. Gitelman (Ed.), "Raw Data" is an Oxymoron (pp. 147–166). Cambridge: MIT Press.

Rosenberg, D. (2013). Data Before the Fact. In "Raw Data" is an Oxymoron (pp. 15–40). Cambridge: MIT Press.

Ruppert, E. (2013). Doing the Transparent State : open government data as performance indicators. In J. Mugler & P. S.-J. (Eds.), A World of Indicators: The production of knowledge and justice in an interconnected world (pp. 51–78). Cambridge: Cambridge University Press.

Shapin, S. (1989). The Invisible Technician. American Scientist, 77, 554–563.

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. Information Systems Research, 7(1), 111–134.

Strathern, M. (Ed.). (2000). Audit Cultures. New York: Routledge.

Walford, A. (2013). Transforming Data: An Ethnography of Scientific Data from the Brazilian Amazon. PhD Thesis, IT University of Copenhagen.