HAL
open science

“Brutification” et instauration des données. La fabrique attentionnée de l’open data

Jérôme Denis, Samuel Goëta

To cite this version:

Jérôme Denis, Samuel Goëta. “Brutification” et instauration des données. La fabrique attentionnée de l’open data. 2016. hal-01347301

HAL Id: hal-01347301

https://minesparis-psl.hal.science/hal-01347301v1

Preprint submitted on 20 Jul 2016

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.

institut interdisciplinaire de l'innovation
WORKING PAPERS
Document de travail i3

"Brutification" et instauration des données

LA FABRIQUE ATTENTIONNEE DE L'OPEN DATA

Jérôme Denis
CSI
I3 (CNRS UMR 9217) – Mines ParisTech
jerome.denis(a)mines-paristech.fr

Samuel Goëta
SES
Telecom ParisTech
samuel.goeta(a)telecom-paristech.fr

Working Paper 16-CSI-01
Juillet 2016

Pour citer ce papier / How to cite this paper: Denis, J. & Goëta S. (2016) "Brutification" et instauration des données. La fabrique attentionnée de l'open data. i3 Working Papers Series, 16-CSI-01.

L'institut interdisciplinaire de l'innovation

(UMR 9217) a été créé en 2012. Il rassemble :

- les équipes de recherche de MINES ParisTech en économie (CERNA), gestion (CGS) et sociologie (CSI),
- celles du Département Sciences Economiques et Sociales (DSES) de Télécom ParisTech,
- ainsi que le Centre de recherche en gestion (CRG) de l'École polytechnique,

soit plus de 200 personnes dont une soixantaine d'enseignants chercheurs permanents.

L'institut développe une recherche de haut niveau conciliant excellence académique et pertinence pour les utilisateurs de recherche. Par ses activités de recherche et de formation, i3 participe à relever les grands défis de l'heure : la diffusion des technologies de l'information, la santé, l'innovation, l'énergie et le développement durable. Ces activités s'organisent autour de quatre axes :

- Transformations de l'entreprise innovante
- Théories et modèles de la conception
- Régulations de l'innovation
- Usages, participation et démocratisation de l'innovation

Pour plus d'information : http://www.i-3.fr/

Ce document de travail est destiné à stimuler la discussion au sein de la communauté scientifique et avec les utilisateurs de la recherche. Son contenu est susceptible d'avoir été soumis pour publication dans une revue académique. Il a été examiné par au moins un referee interne avant d'être publié. Les considérations exprimées dans ce document sont celles de leurs auteurs et ne sont pas forcément partagées par leurs institutions de rattachement ou les organismes qui ont financé la recherche.

The Interdisciplinary Institute of Innovation

(UMR 9217) was founded in 2012. It brings together:

- the MINES ParisTech economics, management and sociology research teams (from the CERNA, CGS and CSI),
- those of the Department of Economics and Social Science (DSES) at Télécom ParisTech,
- and the Management Research Center (CRG) at Ecole Polytechnique,

meaning more than 200 people, including 60 permanent academic researchers.

i3 develops a high-level research, combining academic excellence and relevance for the end users of research. Through its teaching and research activities, i3 takes an active part in addressing the main current challenges: the diffusion of communication technologies, health, innovation, energy and sustainable development. These activities are organized around four main topics:

- Transformations of innovating firms
- Theories and models of design
- Regulations of innovation
- Uses, participation and democratization of innovation

For more information: http://www.i-3.fr/

This working paper is intended to stimulate discussion within the research community and among research users. Its content may have been submitted for publication in academic journals. It has been reviewed by at least one internal referee before publication. The views expressed in this paper are those of the author(s) and not necessarily those of the host institutions or funders.

ABSTRACT :

Alors que les politiques d'open data se multiplient dans le monde, on ne sait que peu de choses des conditions concrètes de l'ouverture des données publiques. En s'appuyant sur une enquête de deux ans menée dans plusieurs institutions françaises, cet article met en lumière les opérations qui sont accomplies en amont de la diffusion des données. Il montre que dans un premier temps, les services passent par une véritable enquête collective pour identifier des jeux de données. Une partie de ces données est ensuite extraite des systèmes d'information qui les hébergent, avec plus ou moins de difficultés. Enfin, les données sont elles-mêmes directement transformées : elles sont nettoyées et subissent des modifications qui leur assurent une double intelligibilité, humaine et technique. Au regard de ces interventions on comprend qu'une grande partie des données ne préexiste pas aux projets d'open data et que leur ouverture est un processus d'instauration qui les fait progressivement advenir. Tandis qu'on pourrait être tenté d'opposer cette fabrique laborieuse des données aux exigences des promoteurs de l'open data, qui insistent sur la nécessité de diffuser des données « brutes », l'article montre qu'elle peut être au contraire appréhendée comme un travail de brutification, un façonnage attentionné de données brutes. Cette description du processus d'ouverture des données publiques invite finalement à reconsidérer ce que l'on entend par « donnée » en général. L'enquête montre en effet que ce qui compte comme données fait l'objet d'explorations et de négociations, ce qui plaide pour l'adoption d'une définition qui en reconnaît le caractère relationnel et transactionnel des données.

KEYWORDS :

open data, données brutes, travail, formats, intelligibilité

En 2013, les chefs d'État qui participaient à la réunion du G8 de Lough-Erne ont signé une charte qui, dans son premier principe, établit que l'ouverture des données publiques doit devenir la pratique par défaut de leurs administrations. Même si elle n'engage qu'assez faiblement les états qui l'ont ratifiée, cette charte apparaît comme l'aboutissement — provisoire — d'un long processus qui a amené les données publiques à devenir des objets politiques à part entière, dont la diffusion et la circulation sont aujourd'hui considérées comme l'un des moteurs principaux de la transparence, de l'innovation, voire de la démocratie. Depuis quelques années, la diffusion des données publiques a été établie comme une obligation légale dans de nombreux pays. Des centaines de portails en ligne sont accessibles, sur lesquels on peut trouver des jeux de données extrêmement variés, mis à disposition par des gouvernements, des municipalités, des institutions et parfois des entreprises.[1]

Cette place inédite des données dans la vie publique des démocraties contemporaines a déjà fait l'objet de nombreuses études. Plusieurs chercheurs, issus de domaines disciplinaires différents, se sont notamment attachés à comprendre les particularités techniques, politiques, mais aussi cognitives de ce que certains considèrent comme un régime de transparence « data-driven ». Du point de vue du « contenu » d'abord, les analyses critiques se sont multipliées, qui s'inquiètent du type de connaissance que les données, réputées parler d'elles-mêmes, produisent. Dans la lignée de Scott (1998), certains ont par exemple insisté sur le fait qu'une transparence basée exclusivement sur des phénomènes quantifiés donne à voir un monde statique et largement schématique (Donovan, 2012). Parce qu'elles sont considérées comme objectives par essence (Birchall, 2014), les données naturalisent des points de vue, et font advenir des réalités au sein desquelles la diversité informationnelle est réduite (Johnson, 2013), laissant peu de place aux savoirs tacites ou peu formalisés d'un grand nombre de collectifs défavorisés (Raman, 2012). On retrouve dans ces critiques une bonne partie des analyses qui ont émergé avec les *alternative accounting studies*, qui ont remis en question l'apparente neutralité des outils de calcul dans les organisations privées et publiques, et ont plus généralement remis en question le réalisme des théories comptables (Rose, 1991 ; Vollmer, 2009). Carruthers et Espeland (1991) à propos du livre de compte à double entrée, ou Miller et O'Leary (1987) à propos de la théorie standard des coûts ont par exemple montré que les dimensions scientifiques et techniques des dispositifs comptables étaient toujours éminemment politiques. Les comptes-rendus quantifiés configurent une certaine réalité dont les fondements mêmes et les conditions de production restent invisibles et indiscutables.

[1] Depuis le lancement de data.gov en 2008 et de data.gov.uk en 2010, le modèle de la « data-driven transparency » s'est exporté, parfois sous la forme de packages clés-en-main proposés aux pays (Birchall, 2015). À l'heure où nous écrivons cet article, plus de 50 pays ont mis en place une politique d'open data selon la dernière édition de l'Open Data Barometer de la World Wide Web Foundation.

Cette perspective invite à critiquer le déterminisme technique des initiatives d'open data. Comme Yu et Robinson (2012) ou Morozov (2014) l'ont montré, le mouvement qui a mené à installer l'open data dans l'agenda politique s'est appuyé principalement sur des questions techniques, laissant largement sous silence ses propres postures politiques. La plupart des appels à la diffusion généralisée de données, tout comme la plupart des projets opérationnels d'open data, sont guidés par un optimisme technologique sans bornes, et présentent l'ouverture des données comme le moyen mécanique d'un élargissement de la transparence publique et d'un *empowerment* des citoyens. C'est cette neutralité de la transparence par les données qui est remise en question, ne serait-ce que parce qu'elle a des conséquences importantes sur la manière dont la vie privée elle-même est définie et renégociée dans les démocraties contemporaines (Meijer, 2009). La focalisation sur les données défavorise des formes alternatives de dévoilement (Birchall, 2014) et ne renforce le pouvoir d'agir que de ceux qui sont déjà « empowered » (Gurstein, 2011 ; McClean, 2011). Plus généralement, parce qu'elles sont largement centrées sur cette transparence technique, les politiques d'open data peuvent également être appréhendées comme une nouvelle étape dans l'avènement des « cultures de l'audit » (Power, 1997 ; Strathern, 2000), une étape qui transforme radicalement le mode d'existence de l'état et de ses publics (Ruppert, 2013).

Ces études critiques offrent de précieuses ressources pour questionner l'ambition des initiatives d'open data et comprendre leurs possibles conséquences politiques. La plupart d'entre elles se positionnent toutefois à un niveau de discussion très général et ne se penchent pas sur les conditions concrètes de cette *data-driven transparency* (Hansen & Flyverbom, 2014). On ne sait pas grand-chose des pratiques situées par lesquelles les données parviennent à circuler, au-delà de la prétention générale à l'immédiateté qui caractérise leurs usages (Mazzarella, 2006). Pour utiliser les termes de Woolgar et Neyland, il manque à ce regard foucaldien porté sur la gouvernance appréhendée par ses dispositifs et ses théories un regard complémentaire sur la « situated and mundane governance » (Woolgar & Neyland, 2013) des données. Les conditions concrètes de l'ouverture des données publiques restent largement inexplorées. Or, la « libration » des jeux de données publiques n'a rien d'une opération anodine. Au sein des administrations, des équipes se sont peu à peu organisées, des départements ont été réagencés pour réaliser les interventions qui permettent l'ouverture. Comment ces équipes travaillent-elles ? Quelles sont exactement les interventions qu'elles effectuent en coulisse des portails d'open data ? Et qu'arrive-t-il aux données « elles-mêmes » ? Par quels états et quelles transformations passent-elles pour être ouvertes ?

L'objectif de cet article est de faire un premier pas vers la réponse à ces questions en explorant les conditions dans lesquelles une série d'initiatives d'ouverture des données publiques ont été réalisées en France. Pour cela, nous nous appuierons sur une enquête ethnographique de deux ans menées au sein d'organisations de l'administration française

qui ont mis en œuvres des projets d'open data (les villes de Paris, Rennes, Poitiers et Montpellier, Etalab, ainsi qu'une grande entreprise de service). Cette enquête a pris la forme d'une série d'observations directes de réunions, et d'entretiens approfondis. Elle a également été l'occasion de récolter un grand nombre de documents, publics et internes.

En les exploitant partiellement ici, nous voulons partir d'un aspect bien particulier de l'open data qui a été largement négligé dans les quelques travaux de recherche qui se sont penchés sur cet objet jusqu'ici : l'insistance, si ce n'est l'obsession, pour le caractère *brut* des données qui doivent être diffusées. Au cœur des grands principes qui sont exposés depuis les premières initiatives d'open data à travers le monde, est en effet rappelée, sous des formes variées, la nécessité de mettre à la disposition du plus grand nombre non pas des données « en général », mais des données « non modifiées », « inaltérées », ou encore « primaires. » Que signifie concrètement cette insistance ? De quoi parle-t-on lorsque l'on parle de données brutes ? Comment fait-on du côté des services concernés pour s'atteler à la mise en circulation de ce genre de données ?

Dans un premier temps, nous proposons de revenir rapidement sur cet aspect afin d'essayer de comprendre ce qu'implique cette focalisation de la part des promoteurs et des instigateurs des politiques d'open data sur les données brutes. Nous exposerons ensuite les résultats de notre enquête en mettant en lumière deux points essentiels qui caractérisent le parcours des données dans le processus qui mène à leur ouverture. Nous montrerons d'abord que l'identification même des données n'est pas une affaire simple, et qu'elle implique, des premiers repérages jusqu'à l'extraction technique de fichiers numériques spécifiques, des explorations collectives, des discussions, et une série d'opérations qui instaurent progressivement des informations aux contours flous en *données*. Dans un second temps, nous montrerons que ce mouvement passe aussi par des transformations qui donnent à voir une dimension importante du processus d'ouverture : la mise en œuvre de l'intelligibilité des données. « Nettoyées », les données sont également adaptées pour être compréhensibles par le plus grand nombre, mais aussi formatées pour devenir *machine readable*. C'est donc un véritable travail des données qui est accompli dans les coulisses des politiques d'open data, une série de délicates opérations qui présentent chacune des risques et un coût sociotechniques. Nous insisterons ensuite sur les conséquences organisationnelles de la mise en œuvre de ces opérations. Celles-ci ne touchent pas en effet simplement les jeux de données en tant que tels, mais ont des répercussions sur l'organisation même du travail au sein des administrations. Enfin, nous reviendrons sur les implications de la mise en lumière du travail des données quant à l'idée même de « donnée brute. » En nous inspirant du vocabulaire de l'une des personnes que nous avons interrogées, nous verrons que, loin d'être contradictoires, les deux dimensions (« brutes » et « travaillées ») méritent d'être pensées ensemble, les opérations décrites pouvant être

comprises comme un processus de *brutification* par lequel les données publiques ouvertes sont façonnées de manière attentionnée.

« We Want Raw Data! »

Revenons à la charte du G8 pour comprendre un peu plus précisément les enjeux des politiques d'open data. On trouve dans le deuxième principe qu'elle expose une phrase qui détaille le type de données que les gouvernements qui la signent s'engagent à mettre à la disposition du public :

> Principe n°2 : De qualité et en quantité
>
> Nous :
>
> - diffuserons des données ouvertes de grande qualité qui soient à jour, complètes et exactes. Dans la mesure du possible, les données seront disponibles sous leur forme initiale non modifiée, et présenteront le meilleur degré de granularité possible.
>
> *http://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/charte-g8-ouverture-donnees-publiques-fr.pdf*

Ce ne sont pas n'importe quelles données qui sont concernées : celles qui comptent sont les données « non modifiées. » Cet intérêt pour les données brutes issues des administrations est relativement récent. Le premier appel explicite pour leur diffusion a été formulé pour la première fois à l'occasion de la rencontre de Sebastopol (États-Unis) en 2007, à l'issue de laquelle une série de principes fondateurs pour l'ouverture des données publiques a été publiée.[2] Le texte qui énonce le deuxième de ces principes souligne l'importance de l'accès à des données qui n'ont pas été traitées, appelées ici *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. » Cette exigence a été réitérée par Rufus Pollock, économiste fondateur de l'*Open Knowledge Foundation*, qui écrivit dans un billet de son blog : « give us the data raw, give us the data now ». L'expression est devenue célèbre deux ans plus tard lorsque Tim Berners-Lee, l'inventeur du web, futur responsable de la politique d'open data du Royaume-Uni, a demandé au public d'une de ses conférences TED de crier en cœur « we want raw data! »

Ce plaidoyer pour la mise en circulation de données brutes, devenue l'une des conditions essentielles de toute politique d'open data, dessine une image très particulière de l'information administrative. Avant toute chose, il constitue les données elles-mêmes en

[2] "8 Principles of Open Government Data" https://www.opengovdata.org

entité informationnelle à part entière. C'est en fait un élément jusque-là inconnu dans le paysage de l'information publique qui est évoqué dans les discours de Pollock et Berners-Lee, dans les principes publiés après la conférence de Sebastopol ou encore dans la charte du G8. Ces appels à l'ouverture des données ne parlent pas des « dossiers » archivés dont la mise à disposition du public est au centre des démocraties modernes depuis la Révolution française (Kafka, 2012). Ils ne concernent pas non plus les statistiques publiques, dont la production et la circulation sont elles-mêmes au cœur de la gouvernementalité contemporaine (Desrosières, 1993). Les « données » dont il est question, même s'il est impossible d'en trouver une définition précise, semblent désigner un type d'information à part, qui existerait en quelque sorte « avant » les dossiers, et en amont des statistiques.

Mais les différents appels à l'ouverture des données n'agissent pas seulement comme des désignateurs d'objets informationnels jusque là ignorés. Tous font un postulat plus générique encore. Ils font l'hypothèse jamais questionnée que l'on trouve dans la plupart des administrations ce type d'objets — des données à l'état pur qui pourraient être transmises à tous ceux qui pourraient en avoir besoin — en grande quantité. C'est ce dont témoigne l'usage massif de la métaphore de la ressource naturelle pour décrire les données publiques. Les administrations seraient « assises » sur une vaste réserve de données dormantes qui ne représenteraient rien de moins que le « pétrole du XXIe siècle.[3] »

Plus encore, cette insistance sur les données brutes laisse aussi entendre que l'ouverture est en soi un processus relativement simple. Toute modification étant bannie, tout « raffinement » du pétrole étant considéré comme une altération, les initiatives d'open data devraient se résumer à une opération quasiment mécanique de mise à disposition. On retrouve dans ce postulat les racines que les mouvements de lutte contre la propriété intellectuelle partagent avec la cybernétique et la communauté du logiciel libre et qui peuvent se résumer en un slogan : « information wants to be free » (Turner, 2006 ; Kelty, 2008 ; Johns, 2009). Les données publiques sont non seulement présentes massivement dans les administrations, mais elles sont également vouées à circuler naturellement, sans entraves et sans manipulations, afin d'être reprises telles quelles par toutes sortes d'usagers potentiels.

Ces postulats et ces prétentions ne sont pas sans rappeler des éléments bien connus des *Science and Technology Studies*. En effet, cela fait déjà longtemps qu'il est question de données brutes en science et que les chercheurs ont souligné l'ambigüité du terme qui, dès que l'on prend en considération le travail quotidien que demande la circulation efficace de n'importe quel jeu de données d'un laboratoire à un autre, voire d'une discipline à une autre, apparaît comme un oxymoron (Bowker, 2000 ; Gitelman, 2013). Dans les faits, et

[3] Expression que l'on pouvait trouver dans *Wired* en juillet 2014, parmi tant d'autres publications : http://www.wired.com/insights/2014/07/data-new-oil-digital-economy/

malgré la mise en place d'infrastructures toujours plus performantes, les données ne circulent jamais de manière parfaitement fluide. Ce point a notamment été mis en lumière par Paul Edwards et ses collègues, qui ont analysé les modalités concrètes de partage et d'usage de données dans des projets de collaboration scientifique à grande échelle (Edwards, 1999 ; Edwards, 2010 ; Edwards *et al.*, 2011 ; Edwards *et al.*, 2013).

Edwards utilise la notion de « frictions » pour décrire les difficultés que représentent la circulation et le partage de données dans différentes disciplines scientifiques. En partant notamment de l'histoire de la climatologie, il souligne par ce terme le coût qu'implique la réutilisation de données qui ont été produites dans des configurations techniques et disciplinaires hétérogènes. Pour alimenter les modèles qui visent à mesurer le réchauffement climatique, il faut par exemple rassembler des enregistrements qui ont été effectués dans des lieux et des temps distincts, mais aussi à des fins et avec des moyens extrêmement variés. Plus on remonte dans le temps et moins ces données sont commensurables : les unités de mesure diffèrent, les instruments ne sont pas calibrés de la même manière (les degrés de précision ont évolué au fil des décennies, les conditions mêmes de la mesure ont changé, etc.). À cela s'ajoutent les différences géographiques, institutionnelles, etc. Edwards explique que ces données ne portent pas en elles les qualités suffisantes pour être utilisées par les scientifiques en question. La récolte des données nécessaires à la mise en calcul d'un climat global, à l'échelle de la terre entière, n'est donc pas un processus transparent qui se résumerait à un échange de flux d'information brute. La circulation des données génère irrémédiablement des frictions.

> Climatology requires long-term data from many locations, consistent across both space and time. This requirement implies a lengthy chain of operations, including observation, recording, collection, transmission, quality control, reconciliation, storage, cataloguing, and access. Every link in this chain represents an information interface subject to data friction. Every point at which data are moved or transformed represents an opportunity for data loss or corruption. (Edwards, 2010, p. 84)

Pour minimiser ces frictions, et pour que les données soient véritablement utilisables, une série d'opérations sont mises en œuvre, qui assurent progressivement leur solidité, leur consistance, leur cohérence, et leur sens même. En d'autres termes, les données ne sont pas seulement partagées, elles sont reconfigurées. Toute forme de collaboration scientifique qui s'appuie sur un échange de données implique un travail complexe, qui passe non seulement par des paramétrages et des réglages à même les données, mais aussi par des interactions directes avec leurs producteurs initiaux. Courriers électroniques, coups de téléphone, réunions sont des ressources essentielles pour clarifier le statut de tel ou tel jeu de données, autour de questions qui peuvent paraître triviales, mais qui sont vitales à la réussite du projet (Edwards *et al.*, 2011). Ces observations nous amènent bien loin d'un idéal de circulation

transparente où les données brutes passeraient d'un service à l'autre à l'intérieur et à l'extérieur des laboratoires scientifiques.

Mais qu'en est-il des données publiques ? Par quelles opérations passent-elles pour circuler ? Si l'on peut trouver des similitudes importantes avec les projets scientifiques étudiés par les *Information Infrastructure Studies*, certaines différences sont également flagrantes. L'open data ne s'inscrit pas, par exemple, dans des programmes d'échange aux interlocuteurs clairement identifiés. La libération des données dessine un horizon très large d'usagers potentiels sans que puisse s'élaborer un protocole bipartite strict. Une autre différence tient au statut des données au sens très général du terme. Le vocabulaire de la donnée est une quasi-évidence en sciences. Aussi bien les chercheurs que les techniciens savent bien qu'ils travaillent avec ce que l'on appelle (et qu'ils appellent) des « données. » Ça n'est pas le cas des administrations, au sein desquelles parler de données n'a rien d'évident.

Cette absence d'évidence représente à nos yeux une opportunité. Enquêter sur l'open data revient en effet à explorer des situations où non seulement la circulation des données pose des problèmes, mais où la définition même de ce qu'est une donnée, de ce qu'elle pourrait être, ou de ce qu'elle devrait être, est instable. En d'autres termes, les projets d'open data offrent un site de recherche fertile pour comprendre comment l'ontologie même des données est explicitement, et parfois difficilement, *énactée*, pour reprendre les termes que Woolgar et Neyland (2013) empruntent à Smith (1974). En conséquence, l'enquête sur l'open data invite à déplacer sensiblement la perspective qui a été majoritairement adoptée pour étudier le partage des données en sciences. Les recherches dans ce domaine se sont en effet essentiellement focalisées sur les opérations qui permettent, en aval, la circulation des données d'un monde social à l'autre. C'est le déplacement des données et ses conséquences qui sont au cœur des problématiques de ces travaux. Comprendre les processus d'ouverture des données publiques suppose de se positionner en amont de leur mise en circulation, et donc d'interroger les conditions concrètes de leur production. C'est ce que nous proposons de faire en reconstituant les principales étapes des projets d'open data que nous avons pu étudier.[4]

De l'identification à l'extraction : l'instauration des données

Suivre les politiques d'ouverture des données publiques au plus près des services directement concernés permet de comprendre l'importance d'une première opération : l'identification des jeux de données qui sont voués à être ouverts. Les premiers travaux qui

[4] Notre objectif n'est pas de décrire chaque cas dans toutes ces spécificités, mais au contraire d'assembler en les stylisant, des moments clefs de l'ouverture des données qui permettent d'en saisir la richesse et les difficultés.

ont discuté des limites de l'open data ont déjà montré à quel point l'étape de la sélection des données était délicate (McClean, 2011 ; Donovan, 2012 ; Raman, 2012 ; Johnson, 2013). Mais ils ont surtout montré qu'elle était *politiquement* importante. La décision de rendre publiques ou non certaines données définit en effet le cadre même de la transparence. Elle en délimite le périmètre, en focalisant l'attention sur tel ou tel aspect de la réalité, tout en laissant dans l'ombre d'autres. Mais cette lecture politique, qui appréhende bien souvent l'open data dans une opposition binaire entre secret et transparence, entre fermeture et ouverture, réduit l'étape de l'identification des données à un processus de sélection. En d'autres termes, elle suppose l'existence de données clairement identifiables et critique les conditions de leur échantillonnage, considéré comme point névralgique de la fabrique de la transparence.

Il y a pourtant une autre identification avant cette identification. Et même si elle peut apparaître trop « technique », il nous semble qu'il est important de comprendre la mise en œuvre de cette étape, notamment parce qu'elle est largement invisible et que cette invisibilité renforce l'évidence — partagée par les promoteurs de l'open data — de la présence de données « déjà-là. » Comme nous l'évoquions plus haut, contrairement à de nombreuses activités scientifiques qui sont entièrement centrées sur la production ou le traitement de données, rares sont les pratiques administratives qui font des données une entité informationnelle appréhendée explicitement. Une grande majorité des agents ne considèrent pas qu'ils ont affaire à des données, encore moins qu'ils en produisent. Dans de nombreux services que nous avons étudiés, ce qui était désigné comme de potentielles données à l'issue des premières discussions avait généralement été jusque-là appréhendé comme des instruments opérationnels, des outils de l'activité administrative. C'est que l'information dans le quotidien de l'administration ne prend que rarement la forme cristallisée de « commodités » dont chacun connaîtrait les frontières et les conditions de circulation. Ainsi, au fil de notre enquête, nous avons découvert que trois points s'avéraient particulièrement problématiques au démarrage des projets open data : la nature des données présentes dans l'institution, leur emplacement, mais aussi, et surtout leur existence même. Avant de se poser la question « comment allons-nous ouvrir tel ou tel jeu de données ? », les agents chargés de mener à bien un projet d'open data se posent donc une question plus abyssale encore : « de quelles données disposons-nous ? »

L'identification des données n'est pas ici une affaire de sélection, mais d'exploration. Une opération cruciale et complexe qui ne peut se réduire à la récolte d'entités clairement définies. Elle s'effectue progressivement, au fil d'échanges avec les services internes, dans un processus incrémental. Elle fait émerger de nouvelles pistes et 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 voie donc, en gros, c'est un peu exponentiel. (Chef de projet open data d'une collectivité locale)

La « récolte » consiste donc en réalité en une enquête incertaine, au fil de laquelle l'identification des données se stabilise peu à peu. Cette enquête est collective et passe par de nombreuses discussions au sein des services. Dans les administrations que nous avons étudiées, ces discussions et les sollicitations des services ont pris des formes variées selon les situations : 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.

On le voit, l'inventaire des données n'a rien de mécanique. Il s'appuie sur des réunions, des discussions, et parfois même de véritables négociations qui traitent conjointement de 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 certaines données, les difficultés ou facilités techniques de leur diffusion, etc. Non seulement les données ne sont pas disponibles de manière évidente dans les services, mais elles ne sont pas non plus sélectionnées en suivant des critères simples qui auraient été définis à l'avance. Au fil de l'exploration collective, les données sont au contraire graduellement co-construites. Ce processus, et l'enquête collective sur laquelle il s'appuie donnent à voir l'identification comme une *instauration* technique, organisationnelle et politique.[5]

> Nous, on explique aux services qu'on s'intéresse à des données, c'est-à-dire soit des fichiers métiers, soit des extractions de bases de données. On part vraiment de la compréhension de ce qu'ils produisent et on leur dit « voilà, la donnée pour nous c'est ça. » […] Typiquement, des gens vont nous présenter des trucs, moi je bosse avec ça, bah y'a des chiffres. Ils vont te présenter une brochure papier ou un rapport d'activité annuel. Il y a des chiffres, quelques tableaux qui se baladent, un ou deux camemberts. Nous on y va, on leur dit « si ça peut être de la donnée. » Si on reprend les chiffres, au final, on va avoir un tableau de cinquante lignes et ça fera sens. (Chef de projet open data d'une collectivité locale)

[5] Latour (2010) emprunte le terme d'instauration à Étienne Souriau pour remettre au cœur de la description des technologies les idées de construction et d'émergence, tout en se débarrassant des défauts métaphoriques du constructivisme. Ici, il nous permet d'insister à la fois sur le fait que les données ne préexistent pas au processus de leur ouverture, mais que celui-ci n'est en revanche pas arbitraire, qu'il rencontre des résistances techniques, politiques, organisationnelles, et prend des tournures inattendues pour celles et ceux qui le mettent en œuvre.

Les données ne sont pas révélées comme ouvrables, découvertes parmi une masse d'autres données 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 identifier : 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 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. C'est ce que l'on comprend, en miroir, des documents qui ont circulé lors de la mise en place du projet, dans lesquels les informations qui ne faisaient pas déjà l'objet d'une diffusion publique étaient purement et simplement qualifiées de *non-data*.

Ainsi, qu'il engage une exploration au long cours, ou s'opère par un repérage plus facile, le processus d'identification est toujours génératif. Il fait advenir une certaine réalité (Law, 2009), un périmètre de données qui ne sont pas simplement désignées comme ouvertes (ouvrables, au départ), mais comme « données » tout court.

Toutefois, le simple fait que les données candidates à l'ouverture soient identifiées ne veut pas dire que les personnes qui sont en charge de la politique d'open data peuvent véritablement y accéder. En réalité, mettre la main sur ces données n'est pas toujours une mince affaire, certaines d'entre elles étant « enfermées » dans des bases de données. Leur ouverture suppose qu'elles puissent être extraites des outils qui jusque-là permettaient leur consultation. Les bases de données relationnelles donnent en effet accès aux données par l'intermédiaire d'interfaces dédiées, appelées « vues utilisateurs » (Codd, 1970). Destinées à simplifier drastiquement la manipulation des données, ces vues supportent une variété d'usages, tout en évitant aux usagers d'avoir à se préoccuper de l'organisation physique des données (Dagiral & Peerbaye, 2013 ; Castelle, 2013). Libérés de ces difficultés, les usagers se trouvent dépendants de ces vues qui sont le seul moyen pour eux d'approcher les données. Et rares sont les bases de données qui proposent des fonctionnalités d'export qui permettraient de récolter les jeux de données sans passer par leur interface de visualisation.

> 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 à l'identification des données, c'est donc une autre exploration qui démarre, à travers laquelle les gestionnaires de bases de données cherchent à récupérer les données « elles-mêmes ». Pour cela, ils doivent en comprendre les modalités de stockage et d'organisation, plonger au-delà des interfaces de visualisation, dans les entrailles des disques durs, à la racine des bases de données.

> Il faut bien comprendre c'est qu'au départ pour la plupart des systèmes, des applications qu'on a chez nous, qu'on a achetés, ils ne sont pas du tout conçus 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)

Les techniciens chargés de cette extraction cherchent à passer au-delà des « vues usagers » pour découvrir la vue physique de la base de données. Mais cette vue physique, lorsqu'elle est finalement reconstituée, n'est pas universelle. À chaque base de données son système d'organisation matérielle des données.

> 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)

Ce sont donc généralement des outils *ad hoc* qui sont développés pour l'extraction. Cette personnalisation des instruments est d'autant plus difficile que les bases de données, les logiciels qui y donnent accès et leurs versions se sont accumulés au sein des organisations. C'est parfois une véritable foule de dispositifs attachés aux données à laquelle les informaticiens font face, dont l'exploration représente un coût 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 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)

Ces opérations d'extraction constituent une deuxième étape dans l'instauration progressive des données. Les bricolages mis en œuvre pour l'extraction, les explorations et les « moulinettes » qui permettent d'atteindre les données, offrent une idée de l'épaisseur de la nasse sociotechnique dont celles-ci doivent être littéralement extirpées. D'autant plus que les questions que soulève cette lutte contre les logiciels qui fournissent des vues utilisateurs ne sont jamais purement techniques. Les pratiques d'extraction mettent en effet en jeu les relations commerciales qui lient les départements informatiques 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. 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 selon 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. C'est un aspect essentiel de l'instauration des données, qui offre un contraste saisissant avec la métaphore des ressources dormantes qu'il suffirait de libérer pour exploiter. Cette vision de la donnée comme « commodité » (Ribes & Jackson, 2013) est remise en question par les coûts que représente l'extraction, mais aussi par l'ambigüité des relations qu'entretiennent les administrations avec leurs fournisseurs de bases de données. Ceux-ci sont en effet propriétaires des « chemins » et des modalités d'organisation de leurs

bases. L'inaccessibilité des données est en quelque sorte au cœur de leur modèle économique.

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

Ce qui veut dire qu'une grande part du travail d'extraction revient, pour les institutions, à reprendre le contrôle sur leurs données en désarticulant l'agencement technique, commercial et juridique qui les lient à des 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 discutable toute description qui prétendrait l'isoler ou la singulariser. L'ouverture des données publiques passe par ce type de singularisation, qui est bien entendu momentanée. Dans le processus de leur instauration, les données sont séparées des bases qui assuraient jusque-là leur accessibilité dans les services, afin d'être déplacées et inscrites dans un nouvel agencement, dédié à leur ouverture.

Transformations

Si les tâches étroitement liées d'identification et d'extraction sont cruciales dans l'instauration progressive des données ouvertes, elles ne semblent toutefois pas suffire à la publication des jeux de données qu'elles font émerger. En suivant la mise en œuvre concrète de différents projets d'open data, nous avons pu observer que les données elles-mêmes étaient travaillées. Dans le processus de leur ouverture, elles sont l'objet d'interventions variées qui les transforment de manière plus ou moins radicale. On peut schématiquement distribuer ces interventions en deux grands types : celles qui mettent en œuvre un nettoyage des données, et celles qui sont dédiées à l'amélioration de leur intelligibilité.

Nettoyage

Le vocabulaire du nettoyage est massivement utilisé en sciences, et les opérations qui s'y rapportent ont été évoquées dans de très nombreux travaux de recherche. Les processus qui aboutissent à la publication d'un article scientifique ont par exemple été analysés en ces termes. Isolés et consolidés, les résultats scientifiques publiés sont présentés dans une

version purifiée qui est le résultat d'un effacement des conditions désordonnées de leur production (Gilbert, 1976 ; Latour & Woolgar, 1979 ; Knorr Cetina, 1981 ; Lynch, 1982 ; Law, 1986 ; Myers, 1988). Une forme de nettoyage est aussi en jeu du côté des données elles-mêmes lorsque les scientifiques travaillent à séparer les erreurs et les biais dus aux instruments qu'ils utilisent des véritables traces du réel qu'ils cherchent à mesurer (Latour & Woolgar, 1979 ; Lynch, 1985). Avec l'avènement des grands projets internationaux et interdisciplinaires, qui placent l'échange des données au cœur du fonctionnement de la science (Wouters & Reddy, 2001), les enjeux du nettoyage des données ont pris une toute nouvelle importance (Zimmerman, 2008). Nous l'avons déjà mentionné, Edwards a par exemple montré que le passage d'une discipline à une autre (dans son cas, de la météorologie à la climatologie) reposait sur des tâches de nettoyage spécifiques (Edwards, 2010). Plus récemment, Walford a exploré la richesse de ce type d'opérations dans le cadre d'une ethnographie de travaux scientifiques menés en forêt amazonienne (Walford, 2013).

Dans les projets que nous avons suivis, le nettoyage concernait plusieurs aspects. Il consistait d'abord à identifier puis corriger des erreurs dans les jeux de données : des valeurs qui étaient considérées comme anormales, ou des « trous » dans les fichiers, des absences de données. Le nettoyage prenait aussi la forme d'une harmonisation des données. Comme nous l'avons vu, les bases de données, et donc les jeux de données, existent dans des formats et des versions très différentes au sein d'une même institution. Ils sont par ailleurs généralement manipulés par des personnes qui les produisent et les utilisent de manières très spécifiques. Ainsi, des entités que l'on pourrait imaginer identiques apparaissent dans les bases de données sous la forme d'unités, voire d'identifiants complètement différents. Comme dans le cas des programmes d'échange de données scientifiques à grande échelle (Baker & Millerand, 2009), les politiques d'ouverture des données publiques impliquent de réduire ce genre d'écarts et de construire une cohérence à travers les différences et les redondances de jeux de données hétérogènes.

> Typiquement, sur les jeux de données des élections : entre les derniers fichiers des dernières élections, et puis les vieux trucs, les fichiers n'étaient pas présentés 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 ça pour que tous les fichiers se ressemblent et soient structurés pareil. (Chargé de projet open data dans une intercommunalité)

La volonté de diffuser des données irréprochables, vierges de toute erreur ou redondance, semble un aspect important des projets d'open data. Elle montre que ces derniers représentent une épreuve pour certaines données qui, si elles étaient publiées telles quelles, seraient considérées comme étant de mauvaise qualité, alors même qu'elles sont satisfaisantes du point de vue de leurs usages internes. C'est un problème qui a été

largement discuté par les chercheurs qui se sont penchés sur les « bad records » (Garfinkel & Bittner, 1967) et les « false numbers » (Lampland, 2010). Dans les organisations, les données sont utilisées quotidiennement dans des configurations spécifiques, avec des horizons d'usage très restreints. Ce sont des « données métier » qui ne sont ni vraies, ni « de qualité », ni même pertinentes en tant que telles. Ce que certains pourraient considérer comme un manque de cohérence, ou un trop faible degré de précision apparaît au contraire du point de vue de leurs usages professionnels comme des qualités et des sources d'efficacité. Comme l'ont montré Garfinkel et Bittner, il existe de nombreuses « bonnes raisons organisationnelles » pour que ces données, que d'autres taxeraient d'inutiles, persistent et soient même considérées comme précieuses. La précision, la qualité et la « vérité » des données sont des dimensions indexicales. C'est justement parce que les données sont ancrées dans des pratiques professionnelles particulières que leur « ouverture » représente une épreuve. Leur confrontation potentielle avec des domaines d'activité qui n'ont aucun rapport avec cet ancrage initial risque de résulter dans leur stigmatisation. Des données auxquelles personne n'a jamais rien eu à reprocher peuvent se retrouver qualifiées d'inutiles ou d'impropres à l'usage. Des absences jamais remarquées jusqu'ici pourraient devenir des manquements, des approximations sans importance pourraient être considérées comme des erreurs, et des redondances jugées utiles risqueraient d'être identifiées comme sources de possibles problèmes. Le nettoyage des données et les transformations qu'il implique sont le coût à payer pour éviter ces risques autant que faire se peut, et préparer la circulation des données d'un monde d'usages restreints vers l'horizon grand ouvert d'utilisations encore inconnues.

Le caractère inconnu des usages potentiels est un autre aspect important de l'ouverture des données publiques. Le nettoyage est en effet la première étape d'une instauration de données qui ne sont pas seulement « propres », mais qui sont aussi génériques, universelles. Celles-ci doivent convenir à tous les usages possibles. On s'en doute, cette projection des données vers l'universalité des usages n'est pas simple. Elle s'inscrit dans une exigence explicite qui est au cœur de transformations plus radicales encore : l'intelligibilité des données.

Les deux horizons de l'intelligibilité des données

Le fait que les données doivent être intelligibles est au cœur de nombreux textes qui visent à encadrer les politiques d'open data à travers le monde. On le trouve par exemple en bonne place parmi les engagements exprimés dans charte du G8 que nous évoquions au début de cet article, puisqu'il est au cœur de son deuxième principe.

> (Nous :) - veillerons à ce que l'information contenue dans les données soit rédigée en langage simple et clair, de manière à être comprise par tous (…).

> [Annexe technique de la Charte] Principe n°2 : En qualité et en quantité
>
> 6) Nous nous engageons à publier des données qui soient à la fois d'un haut niveau de qualité et diffusées en grande quantité. Lorsque nous publierons des données, nous nous assurerons de le faire d'une manière qui aide chacun à les obtenir et les réutiliser.

Si elle semble évidente, voire triviale, cette exigence d'intelligibilité est loin d'aller de soi. À vrai dire, quasiment aucun jeu de données issu des administrations que nous avons étudiées ne répond à ce critère s'il ne fait pas l'objet d'importantes transformations. Ils contiennent tous des données métier pleines d'idiosyncrasies, de termes incompréhensibles et autres acronymes obscurs.

Il existe un moyen d'améliorer l'intelligibilité des données sans les modifier directement : la production de métadonnées qui permettent notamment de mettre à disposition des usagers une sorte de dictionnaire pour faciliter la compréhension des données. Mais comme l'ont montré les *Information Infrastructure Studies* qui se sont penchées sur la question (Baker & Bowker, 2007 ; Edwards, 2010), non seulement la fabrication de ces métadonnées est coûteuse, mais elle est par ailleurs vouée à demeurer incomplète. Les métadonnées elles-mêmes ne sont pas universelles par nature, et si les frictions propres à la circulation des données sont nombreuses, celles qu'engendrent les métadonnées apparaissent plus grandes encore (Edwards *et al.*, 2011).

Les métadonnées n'étant certainement pas la solution miracle à la fabrique de l'intelligibilité, celle-ci passe donc par des manipulations et des modifications à même les données. Certains termes sont remplacés par d'autres, des colonnes de tableaux sont déplacées, une partie des informations est même parfois purement et simplement effacée. 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. Utiles dans le travail quotidien, 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. (Gestionnaire de bases de données de transport)

On le voit clairement dans l'extrait d'entretien ci-dessus, comme dans le nettoyage, l'enjeu principal de ces modifications est de passer d'un domaine de significations spécialisé, où les mots, les acronymes ou les abréviations prennent place dans une économie langagière en grande partie stabilisée, à un univers de sens élargi, partagé par le plus grand nombre. Cette opération est délicate. Elle passe par des doutes, des discussions collectives, et est appréhendée par celles et ceux qui la réalisent comme une tentative jamais complètement réussie. Tout le monde est d'accord pour reconnaître qu'il est impossible de produire des données compréhensibles par tous.

Mais la question de l'intelligibilité n'est pas réglée avec ces interventions pour autant. La compréhension des données « par le plus grand monde » n'en est en fait qu'un des aspects. On trouve en effet dans les exigences des promoteurs des politiques d'open data, dans les textes qui les encadrent et dans les indicateurs qui permettent de les évaluer un « principe » essentiel : la nécessité de diffuser des données « machine readable. » C'est par exemple explicite dans la charte du G8 que nous avons déjà évoquée, dont le cinquième chapitre mentionne la diffusion de données qui « puissent être lues en blocs par machine » qui permettent un « accès automatique. » De même, les principes publiés par la *Sunlight Foundation* en 2010 précisent : « information should be stored in widely-used file formats that easily lend themselves to machine processing.[6] »

Les données ouvertes doivent donc être non seulement *humainement*, mais aussi *techniquement* intelligibles. C'est en fait une des conditions de leur circulation fluide et l'un des postulats centraux du plaidoyer pour diffuser des « données brutes. » Les données sont censées peupler les disques durs des différents services de l'administration et pouvoir migrer quasi automatiquement vers d'autres disques durs, sans altération. Or, on l'a vu à propos de l'extraction, les choses sont loin d'être si simples. Même lorsqu'elles sont clairement identifiées, ce qui n'est pas toujours le cas, les données existent dans des formats variés, plus ou moins ouverts, et plus ou moins faciles d'accès. Qui plus est, « puissent être lues par une machine » est un terme très vague qui ne dit pas grand-chose de concret. Comment cette exigence est-elle mise en œuvre pratiquement ? Dans les activités que nous avons observées, l'intelligibilité technique est essentiellement assurée par l'adoption de standards, ou de formats partagés. De fait, ce n'est qu'à partir du moment où les données identifiées, extraites, nettoyées, en partie déjà transformées, adoptent tel ou tel format qu'elles sont considérées comme « véritablement » ouvertes. Mais la mise au format d'un jeu de données n'est jamais une opération évidente, sans conséquence sur les données. C'est un geste délicat qui implique lui-même de nouvelles transformations.

[6] https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

Nous n'allons bien sûr pas entrer dans les détails de cette question des formats, qui est d'autant plus complexe qu'elle est émergente dans le domaine de l'open data. Nous proposons ici de nous arrêter rapidement sur le cas qui paraît le plus simple, afin de souligner l'ampleur des conséquences du formatage, le plus anodin soit-il. Dans la plupart des situations auxquelles nous avons pu avoir accès, ça n'est pas un standard détaillé et rigide qui était visé pour la mise en forme des données, mais un format basique, partagé par la plupart des acteurs du domaine, notamment parce qu'il est considéré comme ouvert et lisible par le plus grand nombre d'instruments : le CSV (comma-separated values).

La production d'un fichier en CSV est accessible à presque tout le monde, et le passage d'un document tableur dans ce format peut s'opérer par la fonction « save as » ou « export » de n'importe quel logiciel dédié. Toutefois, cette opération de formatage n'a rien de transparent, au contraire. Si on n'y prend garde, elle peut corrompre définitivement un jeu de données. Le CSV, comme son nom l'indique, affiche des valeurs séparées par des virgules, sans aucune autre information. L'export dans ce format opère donc une « mise à plat » qui suppose une transformation des données qui, nous l'avons vu, existent au quotidien sous des formes riches et variées. En réalité, une grande partie des opérations que nous avons décrites précédemment sont déjà orientées vers la production de données dont le formatage en CSV s'effectuera sans entrave. Le nettoyage, par exemple, s'attache non seulement à écarter des données qui doivent être considérées comme mauvaises or fausses pour différentes raisons, mais aussi aux cases vides ou encore aux lignes masquées. Ces masquages peuvent être utiles à la consultation quotidienne des données, mais incompatibles avec le format CSV.

D'autres opérations qui sont effectuées pour assurer la bonne marche du formatage interviennent directement sur la nature des documents de travail au sein des services. Les options telles que la fusion de cellules, ou l'utilisation des couleurs pour produire des repères dans l'espace doivent par exemple être abandonnées. C'est loin d'être anodin. Ces fonctionnalités investissent dans des propriétés informationnelles et perceptives de l'espace de l'écran, dont les travaux autour de la cognition distribuée ont montré à quel point elles étaient utiles à l'accomplissement du travail (Norman, 1991 ; Hutchins, 1995). Le passage au format CSV nécessite d'effacer purement et simplement ces informations spatiales (Kirsh, 1995) qu'il est incapable de prendre en compte.

Le cas du General Transit Feed Specification (GTFS) est un autre exemple qui permet de voir à l'œuvre d'autres types de transformations. Format imposé par Google et certaines agences partenaires, le GTFS est rapidement devenu un standard *de facto* dans le domaine des données de transport. Désormais ouvert, c'est-à-dire aux spécifications publiques non soumises à une licence restreignant son usage, il se fonde sur le regroupement de plusieurs fichiers CSV dans un fichier compressé au format zip. Chaque jeu de données est composé d'au moins six fichiers décrivant l'agence de transport fournissant les données, les arrêts, les

lignes, les trajets prévus quotidiennement sur chaque ligne, les horaires de passage à chaque arrêt et les dates de passage des véhicules. Les spécifications du standard définissent ce qui doit figurer dans chaque champ (fig. 1).

stops.txt

File: **Required**

Field Name	Required	Details
stop_id	**Required**	The **stop_id** field contains an ID that uniquely identifies a stop or station. Multiple routes may use the same stop. The **stop_id** is dataset unique.
stop_code	Optional	The **stop_code** field contains short text or a number that uniquely identifies the stop for passengers. Stop codes are often used in phone-based transit information systems or printed on stop signage to make it easier for riders to get a stop schedule or real-time arrival information for a particular stop. The stop_code field should only be used for stop codes that are displayed to passengers. For internal codes, use **stop_id**. This field should be left blank for stops without a code.
stop_name	**Required**	The **stop_name** field contains the name of a stop or station. Please use a name that people will understand in the local and tourist vernacular.

Fig. 1: *Information about stops in GTFS files*[7]

Selon les pratiques et les formes utilisées dans les services, un plus ou moins grand nombre d'ajustements est nécessaire pour passer d'un jeu de données métier à une série de données au format GTFS. Le format impose par exemple qu'une station soit dotée d'un identifiant unique. Cette obligation n'est pas forcément d'usage dans les services administratifs et les agences de transport où elle n'a pas d'utilité propre. Il arrive que les identifiants diffèrent d'une base de données à l'autre. Durant notre enquête, nous avons ainsi rencontré le cas d'une société de transport dont le nom des arrêts variait entre le service en charge de la carte du réseau et celui en charge des horaires du service de bus. Or, seul le croisement de ces deux bases de données permettait de disposer des informations nécessaires pour construire le fichier GTFS. Pour que les données soient conformes aux spécifications du format, il a donc fallu « corriger » les bases de données et adopter des identifiants uniques.

Le passage d'un jeu de données vers un format aussi anodin que le CSV, ou vers un standard plus complexe tel le GTFS, met donc en lumière une nouvelle série de transformations qui s'ajoutent aux opérations que nous avons décrites jusqu'ici. L'ensemble donne à voir la richesse d'un processus qui ressemble moins à une « ouverture » mécanique de données

[7] https://developers.google.com/transit/gtfs/reference#stops_fields

(comme s'il suffisait d'ouvrir un robinet d'où coulerait automatiquement de la donnée), qu'à une délicate production d'artefacts informationnels qui sont progressivement instaurés en *données* ouvertes. Ces transformations reposent sur un travail qui est d'autant plus invisible qu'il semble entrer en complète contradiction avec les exigences évoquées au début de cet article de faire circuler des données primaires, inaltérées. Dans les deux prochaines sections, nous revenons sur ces points en insistant d'abord sur les conséquences organisationnelles sur lesquelles repose, et parfois qu'engendre, ce travail, d'un côté, et sur ce qu'il nous permet de comprendre du caractère brut des données, de l'autre.

Conséquences organisationnelles

Si les tâches que nous avons identifiées montrent que les données sont transformées dans le processus, il faut aussi tenir compte de leurs répercussions organisationnelles.

Prenons le cas de l'identification. La mise en place d'un inventaire est rarement envisagée comme une opération unique, mais plutôt comme les premiers pas d'un processus récurrent de circulation des données, des services qui les produisent et les manipulent jusqu'à ceux qui sont chargés de leur diffusion publique. Un responsable de projet open data dans une collectivité locale explique ainsi qu'à terme les gestionnaires de données au sein de son organisation devront 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 [circule] 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 identifiés, 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 redistribue certaines cartes dans l'organisation. Des rôles nouveaux et des responsabilités inédites sont attribués. Et plus cet inventaire continu est outillé (plus il est inscrit dans des infrastructures techniques stables, jusqu'à prendre la forme d'un logiciel de *workflow* dans certains cas), plus il ancre dans le temps cette transformation organisationnelle et en rigidifie les contours.

Les opérations d'extraction et les nombreuses transformations que nous avons rapidement décrites ici ont elles aussi des conséquences organisationnelles. La production de l'intelligibilité des données implique notamment, comme dans le cas des projets de collaboration scientifique, l'émergence de compétences nouvelles et l'ajustement de la division du travail dans les services. L'écriture de métadonnées, la traduction d'idiosyncrasies et de catégories professionnelles en termes génériques, ou encore l'harmonisation d'identifiants sont autant de tâches nouvelles qui doivent être intégrées dans chaque institution. Une place doit être faite pour les « data guys » (Edwards, 2010).

Au-delà même de ces aspects, certaines dimensions mises en lumière ici peuvent avoir des conséquences plus radicales. Dans certaines institutions, il est en effet envisagé de profiter des bouleversements liés à l'open data pour repenser les processus organisationnels au-delà des seuls métiers de l'informatique. Ces réorganisations sont présentées comme un moyen de réduire l'ampleur du travail des données en l'intégrant dès les activités situées en amont de la diffusion. Cela passe par exemple 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 untel ». 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)

Anticiper les transformations des données en amont dans les services est généralement envisagé comme un moyen de moderniser, ou de rationaliser, l'administration publique.

Mais ces réorganisations ne relèvent pas seulement de la reconnaissance de la valeur du travail sur les données, jusque-là invisible, qui serait désormais inscrit dans les processus internes. Elles modifient ce travail et renversent le mouvement que nous avons décrit, en intégrant les problématiques d'ouverture dans des activités qui jusque-là bénéficiaient de données *ad hoc*. Placer les opérations de transformations ou de nettoyage en amont revient à considérer que les données génériques sont, par nature, de « bonnes données. » Or, ce faisant, on débarrasse les données métier de leur qualité et de leur pertinence indexicales. On installe au cœur des activités quotidiennes de l'administration les tensions identifiées par Garfinkel and Bittner (1967) entre des registres d'usage non alignés, voire incompatibles.

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 reconnaît que l'ouverture des données a 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), qu'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 cas, le caractère générique des données — leur aspect « brut » — est considéré comme un bien en soi, sur lequel il faut aligner les idiosyncrasies professionnelles.

Brutification

Revenons sur ce qui est fait aux données « elles-mêmes. » Comment comprendre le processus de leur instauration et les transformations que nous avons mis en lumière si on les confronte aux principes que nous avons évoqués au début de cet article, en particulier à l'obsession des promoteurs de l'open data pour des données « non modifiées ». Que faire de l'idée même de « données brutes » au regard de ce que nous avons observé ?

De même qu'elles ont élaboré une critique solide de la distinction entre science et bricolage, les S*cience and Technolgy Studies* n'ont eu de cesse de remettre en cause l'opposition empruntée à Lévi-Strauss entre cru (« raw ») et cuit (Lévi-Strauss, 1964) en montrant qu'on ne trouvait jamais de données brutes au sens « d'asociales » dans les laboratoires scientifiques. La réalité est toujours prise dans un environnement sociotechnique qui en assure les conditions d'observabilité. Le façonnage des données (leur « cuisson ») consiste à traiter de ce que les chercheurs considèrent comme des biais, dont

l'existence montre que ce qui est appelé « données brutes » est toujours déjà socialisé. Dans son enquête en Amazonie, A. Walford respécifie la notion de données brutes en sciences en montrant que celles-ci apparaissent dans les mains de ceux qui les collectent, puis de ceux qui les transforment, comme des données en devenir, des entités ambiguës, multiples. Ces données brutes attendent en quelque sorte la série des traitements qui vont assurer la consolidation progressive de résultats scientifiques à travers leur inscription dans un réseau sociotechnique stabilisé (Walford, 2013). Cette inscription passe notamment par des dispositifs de défiance vis-à-vis des instruments qui ont permis de récolter les données. Transformer les données pour les partager consiste alors à les débarrasser du « bruit » et des « artefacts » qui viennent parasiter l'accès à la réalité observée.

On trouve un lien évident entre les projets d'ouverture des données publiques et les projets de mise à disposition des données scientifiques, puisque, nous le verrons, une partie du travail qui est effectué en amont sur les données vise également à leur mise en intelligibilité. L'objectif des *data cleaners* observés par Walford est en effet de « make data real », et d'assurer que les données portent en elles les conditions de leur propre détachement de la situation mesurée ou étudiée. Le nettoyage s'opère ici dans une perspective de partage et permet de passer de données brutes à la fois potentiellement parasitées par des biais et trop génériques, à des données certifiées qui peuvent circuler d'un monde à l'autre, d'une discipline à l'autre.

Cependant, dans les programmes d'open data, le processus est en quelque sorte inversé. Les données destinées à la diffusion publique ont déjà eu une longue vie sociale, elles n'émanent pas d'instruments génératifs dont il faudrait se méfier, mais sont ancrées dans des usages parfois anciens. Elles sont déjà inscrites dans des réseaux sociotechniques qui les stabilisent et les orientent vers des pratiques spécifiques. L'enjeu des tâches qui sont accomplies pour assurer leur ouverture n'est donc pas d'en réduire le périmètre en les purgeant des scories contextuelles et instrumentales de leur fabrication, mais au contraire d'en élargir l'usage possible. Les opérations mises en œuvre pour l'ouverture des données visent à les désenclaver de leurs agencements sociotechniques initiaux. En les débarrassant de la gangue qui faisait d'elles des données *métier*, les travailleurs de l'open data s'attachent à les transformer en données à la fois intelligibles et ambiguës, au sens de plurivoques, ouvertes à une série de nouveaux traitements. L'instauration des données ouvertes consiste donc à transfigurer des données aux usages et aux significations étroites, ciblées, en données universelles. Bien entendu, ce désencastrement des données métier ne veut pas dire que les données puissent exister par elles-mêmes, une fois libérées. L'ouverture opère toujours un réencastrement des données dans un nouveau réseau sociotechnique qui opère ses propres formes de réduction et de clôture : notamment celles qu'opèrent les formats et autres standards techniques.

De ce point de vue, les programmes d'open data sont assez proches de ce qu'a observé P. Edwards avec la climatologie (Edwards, 2010). Le partage et la circulation de données déjà utilisées, ancrées dans des pratiques parfois vieilles de plusieurs dizaines d'années, passe par des opérations de transformation qui assurent non seulement qu'elles pourront être intelligibles à leurs futurs usagers, mais également qu'elles pourront être mobilisées par des instruments spécifiques et associées à de nouvelles données (pour être comparées, additionnées, etc.). Les transformations assurent ici non pas le transport d'une réalité d'un monde sauvage et changeant vers un le monde stabilisé du laboratoire (Latour, 1993), mais le passage d'un monde stabilisé à un autre.

Dans son ethnographie du Fonds Monétaire International, R. Harper (1998) a observé le même type d'enjeux à l'occasion des missions des agents. En effet, au fil des discussions informelles et des réunions plus institutionnelles, ceux-ci sélectionnent et transforment peu à peu les données que les représentants du pays qu'ils visitent leur fournissent, afin d'en faire des ressources fiables pour produire les calculs économiques dont ils ont la charge, et les rapports politiques que le FMI publiera à l'issue de la mission. Harper montre que c'est une « transformation morale » des nombres qui est en jeu dans ce processus (p. 227). Il en est de même pour les opérations qui permettent l'ouverture des données, à ceci près que l'ordre moral qu'elles instaurent est moins orienté vers la justesse des données et des représentations de la réalité que vers leur réutilisabilité.

Plus encore que la remise en question d'une définition naturaliste des données brutes, c'est donc la reconnaissance du travail des données qui importe (Denis, 2011 ; Denis et Pontille, 2012). Les données ne flottent pas dans l'air ni ne tombent du ciel. Elles sont manipulées, fabriquées, transformées, ajustées, etc. En d'autres termes, leur production et leur circulation ne s'opèrent pas sans coûts. Les premières ethnographies de laboratoires et les grandes enquêtes historiques ont insisté sur cet aspect en montrant que les données n'étaient pas passivement récoltées mais qu'elles étaient façonnées (Latour & Woolgar, 1979 ; Knorr Cetina, 1981 ; Shapin & Schaffer, 1985). Ces travaux ont en quelque sorte mis en avant le coût de production des données. Plus récemment, la recherche qui s'est développée autour des infrastructures informationnelles, des grands projets scientifiques internationaux et interdisciplinaires a insisté sur les coûts de partage des données. Ici, nous avons tenté de montrer qu'ils existaient également des coûts spécifiques à l'ouverture des données. Leurs particularités tiennent à l'idée même « d'ouverture » et à l'universalité qui la sous-tend. En sciences, les coûts de partage sont essentiellement dédiés à l'articulation de mondes sociaux distincts. Les jeux de données ont des usagers plus ou moins clairement identifiés et le partage se construit en grande partie dans la négociation des pratiques de chacun, et dans l'orientation des données, leur adaptation aux pratiques des « autres » (Millerand & Baker, 2010). Dans le cas des programmes d'open data, les « autres » ne sont pas facilement identifiables. Il est même parfois considéré qu'ils ne doivent pas l'être, au risque de favoriser

des communautés, d'orienter des jeux de données vers des « intérêts » spécifiques aux dépens de l'intérêt général. D'un certain point de vue, dans le processus d'ouverture, les usagers peuvent être considérés comme des « parasites » (Denis & Pontille, 2014). Les coûts d'ouverture se concentrent en partie sur cette nécessité de (re)produire des données intelligibles en soi, réutilisables universellement, sans la présence d'usagers avec lesquels négocier localement tel ou tel aspect des données en voie d'ouverture.

Quelque chose nous a particulièrement frappés dans notre enquête, tandis que nous découvrions la variété des opérations qui étaient mises en œuvre pour concrètement ouvrir certaines données. Si elles nous étaient parfois décrites comme difficiles, et si leur invisibilité, leur non-reconnaissance, et le manque de moyens dont elles pâtissaient semblaient parfois douloureux, leur mise en lumière ne débouchait jamais sur une remise en cause frontale de l'idée même de donnée brute. Comment comprendre cela ? Comment ne pas voir une opposition très claire entre d'un côté la prétention politique et théorique à faire circuler des données inaltérées et de l'autre la description d'opérations qui semblent précisément modifier profondément les données pour assurer les bonnes conditions de leur ouverture ? C'est la responsable d'un projet open data qui nous a apporté la réponse en nous expliquant dans ses propres termes en quoi consistait le travail qu'elle effectuait sur les données, un travail qu'elle présentait comme la condition même de l'existence de données brutes.

> 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… […] 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 simples possible à utiliser ensuite pour les développeurs. (Chargée de projet open data dans une intercommunalité)

Nous pensons que nous avons tout intérêt à nous inspirer de cette description et, au lieu d'opposer les exigences de données brutes d'un côté et les transformations que nous avons observées de l'autre, à nous pencher sur l'articulation entre les deux, c'est-à-dire à appréhender le processus d'ouverture des données comme un travail de *brutification*. Identifier, extraire, nettoyer les données, façonner leur intelligibilité humaine et technique ne sont pas des interventions qui entrent en contradiction avec l'exigence de données brutes. Au contraire, elles sont un moyen de s'y plier. Considérée dans ces termes, la notion même de « données brutes » est bien un oxymore au sens de Bowker (Bowker, 2000 ; Gitelman, 2013), mais un oxymore situé, un oxymore que les travailleuses et les travailleurs des données se chargent de réduire dans leur activité quotidienne.

Conclusions : « obtenues » et « données », une définition relationnelle

Au-delà de la question des données brutes, nous pensons qu'une posture pragmatiste, qui prend les activités et le vocabulaire des acteurs interrogés au sérieux peut aider à reconsidérer la notion même de *donnée*. Borgman l'écrit dans l'introduction d'un récent ouvrage, les rares définitions de ce qu'est une donnée demeurent en effet vagues et — à ses yeux — peu satisfaisantes.

> The inability to anchor the concept in ways that clarify what are data and are not data in a given situation contributes mightily to the confusion about matters such as data management plans, open data policies and data curation. (Borgman, 2015, p. 28-29)

Il est d'ailleurs frappant qu'on ne trouve nulle part dans les documents que nous avons évoqués jusqu'ici une description stable et claire de ce qui peut être ou ne pas être considéré comme une donnée. Tim Bernes-Lee lui-même, pendant la fameuse conférence TED que nous avons évoquée plus tôt n'est pas très précis lorsqu'il parle de données :

> Turns out that there is still huge unlocked potential. There is still a huge frustration that people have because we haven't got data on the web as data. What do you mean, "data"? What's the difference -- documents, data? Well, documents you read, OK? More or less, you read them, you can follow links from them, and that's it. Data -- you can do all kinds of stuff with a computer. (…) In fact, data is about our lives. You just -- you log on to your social networking site, your favorite one, you say, "This is my friend." Bing! Relationship. Data. You say, "This photograph, it's about -- it depicts this person. " Bing! That's data. Data, data, data. There are data in every aspect of our lives, every aspect of work and pleasure, and it's not just about the number of places where data comes, it's about connecting it together. (Tim Berners-Lee, conférence TED, Février 2009[8])

Faut-il considérer cette absence de définition stable et restreinte comme un problème ? Peut-être pas. Peut-être qu'elle invite au contraire à déplacer la manière même dont on peut comprendre ce que sont les données. À vrai dire, après avoir suivi les personnes qui se chargent concrètement des politiques d'open data, le caractère vague de ce qu'est ou ce que n'est pas une donnée semble tout sauf surprenant. Nous avons bien vu que les données ne préexistaient que rarement dans les administrations et qu'elles ne prenaient jamais la forme prédéterminée d'informations numériques que l'on pourrait mettre en circulation automatiquement. Un premier moyen de respécifier la notion de données en prenant acte de ce que nous avons pu montrer consisterait donc à revenir à l'idée d'instauration et à

[8] La transcription complète du discours de Berner-Lee's talk est accessible à cette adresse : http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

rappeler le jeu de mots que Latour : « Décidément, on ne devrait jamais parler de "données", mais "d'obtenues." » (Latour, 1993, p. 188). C'est un aspect évident de notre enquête : les données sont des *outputs* d'un long processus. Mais cet écart dans l'appréhension des données ne peut représenter qu'un premier mouvement dans l'analyse des politiques d'open data, au risque de jeter le bébé des données avec l'eau du bain de leur instauration. Il nous faut aussi comprendre en quoi et comment les données, dans ce processus, deviennent *données* au sens de « déjà-là » pour ceux qui sont désignés comme leurs futurs utilisateurs.

Rosenberg est revenu en détail sur les liens entre la donnée au sens mathématique du terme et ce qui est donné au sens à la fois de déjà-là et de postulat rhétorique (Rosenberg, 2013). La donnée dans les termes des mathématiques, explique-t-il, a pendant longtemps désigné le matériau de base de l'analyse et du calcul, sans que son rapport à la réalité 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. Étudier les pratiques de l'open data de l'intérieur, comme nous l'avons fait, invite à ne pas négliger cet aspect au seul motif d'insister sur les difficultés à obtenir des données prêtes à être ouvertes. Certes, les données qui sont façonnées de manière attentionnée par ceux qui deviennent leurs travailleurs dans ce processus peuvent être considérées comme un résultat, et donc comme des *obtenues*. Mais ces données-obtenues sont elles-mêmes destinées à se muer en données-données pour leurs futurs usagers : des données qui peuvent être utilisées sans qu'on ait à en questionner la solidité, la validité, ni la pertinence. Pour saisir ce mouvement, ce passage d'obtenues à données, on peut plaider pour l'usage d'un autre jeu de mots et comprendre comment, en particulier dans le cas de l'open data, les données sont parfois des *dons* : des artefacts informationnels soigneusement confectionnés, offerts à la communauté.

La notion d'oxymore utilisée par Bowker et Gitelman reste utile dans cette perspective, puisqu'elle aide à faire remonter à la surface le travail qu'un tel don implique. Parler d'oxymore plutôt que d'illusion ou de mythe, permet de mettre en lumière les tensions que les données comme leurs travailleurs rencontrent dans le processus de leur instauration, tensions dont les coûts ne sont jamais complètement anticipés, lorsqu'ils ne sont pas purement et simplement niés.

Il ne faut toutefois pas tomber dans l'angélisme. Ce don se fait dans des conditions très particulières, conditions qui donnent à voir ce qui est attendu par ceux qui sont censés en bénéficier. Les principes de l'open data, les formats utilisés et les standards qui émergent, tout comme les dispositifs internationaux qui évaluent les politiques d'open data à travers le monde sont autant d'instruments qui participent à la définition restreinte, et aujourd'hui encore mouvante, de ce qui, dans le domaine de l'open data, compte comme des données. C'est aussi le cas des nombreuses prises de position, articles, billets de blogs qui publient des critiques parfois fortes à propos des données qui sont mises à la disposition du public,

stigmatisant le format ou la mauvaise qualité de certaines, voire le « contenu » d'autres. Il arrive que les dons ne soient pas acceptés, jugés indignes de ceux qui sont supposés les recevoir. Par exemple, en juillet 2014, lorsque la Haute Autorité pour la transparence de la vie publique (HATVP) a publié un jeu de données recensant les déclarations d'intérêts des élus, *Regards Citoyens*, un collectif dédié à la diffusion et au partage de l'information publique, a diffusé un texte sur le site web qui pointait les problèmes que représentaient à leurs yeux les documents diffusés, appelant les volontaires à participer à l'élaboration de ce que nous avons appelé ici leur intelligibilité technique.

> Si la Haute Autorité pour la Transparence [HATVP] met à disposition un jeu de données recensant les élus et les déclarations qu'elle contrôle, les informations contenues dans les déclarations d'intérêts ne sont en revanche pas à proprement parler en Open Data : elles n'ont pu être publiées par la HATVP que scannées sous la forme de PDF images rendant l'exploitation de ces informations malaisée au vu du grand nombre d'informations mises en ligne.[9]

Le projet « Bad Data » initié par l'Open Knowledge Foundation relève d'un geste similaire. Sur son site web y est dénoncée publiquement la mauvaise qualité de certains jeux de données diffusés sur des portails d'open data. À propos d'un fichier CSV trouvé sur le portail data.gov.uk, on trouve par exemple cette liste détaillée de défauts :

> The problem is the CSV is so messy only a human could use it! What specifically is wrong?
> - The first column is missing a heading (one guesses this should be "date"?
> - Dates are not of a recognizable format instead being of form: "2006/2007 - 1". One assumes this should be a month or similar (but its not entirely clear if these are months since 13 items in a year!)
> - Percentage sign written into percentage column
> - Large number of trailing blank rows and columns[10]

Des déclarations de ce type dessinent explicitement des frontières entre bonnes et mauvaises données, désignant même parfois directement ce qui est et ce qui n'est pas de la donnée. Au même titre que la définition de principes internationaux et l'élaboration de standards techniques, elles nous permettent de comprendre que les données ne peuvent être désignées comme des objets fixes, identifiables par une série de caractéristiques déterminées une fois pour toutes. Au contraire, elles invitent à adopter ce que Leonelli (2015) appelle un « cadre relationnel » pour appréhender les données. Mais au lieu de se

[9] https://www.regardscitoyens.org/numerisons-les-declarations-dinterets-des-parlementaires/
[10] http://okfnlabs.org/bad-data/ex/tfl-passenger-numbers

focaliser sur la seule capacité de certaines informations à faire preuve, qui est le critère que mobilise Leonelli dans le cas des données scientifiques, on peut défendre une posture relationnelle, et transactionnelle, élargie. En adoptant la perspective de Engeström (1990) à propos des instruments au travail, on peut ainsi ne pas se demander « ce qu'est une donnée », mais chercher à comprendre « quand est une donnée. » Dans ces termes, l'enquête que nous avons présentée ici montre qu'au gré du processus d'ouverture des données, des fichiers, des documents, des chiffres, des textes, des images, ne prennent le statut de données que lorsqu'ils arrivent — dans certaines configurations, sous certaines conditions et selon des critères plus ou moins négociés — à être considérés comme des entités déjà-là, des matériaux de base, par ceux qui, à l'issue de cette transaction, deviennent leurs usagers.

Cette posture relationnelle, parce qu'elle est attentive aux ontologies situées, permet de comprendre qu'un « même » objet informationnel peut-être considéré ou pas comme une donnée selon les cas. Elle permet aussi de mettre en lumière le fait que dans la transaction les « documents » des uns deviennent parfois les données des autres, pour reprendre les termes des sciences de l'information et de la communication. Surtout, elle souligne l'importance des critères, conditions et autres configurations qui assurent le succès de la transaction. La possibilité d'imposer dans la relation certains de ces aspects qui fondent la capacité d'informations variées à devenir des données n'est en aucun cas distribuée de manière égalitaire, et elle est une source de pouvoir considérable. Dans le cas des politiques d'ouvertures des données publiques, ce sont les premiers promoteurs de l'open data, issus des communautés du logiciel libre et de l'*open access*, qui ont progressivement consolidé et installé les principes directeurs sur lesquels se sont appuyés, parfois en les reprenant mot pour mot, les textes officiels. Ce faisant, ils ont encadré l'activité de celles et ceux à qui ils demandaient de diffuser des données. En présupposant l'existence de données brutes dans les administrations, en prônant leur circulation fluide, en définissant les critères de leur qualité, et en exigeant leur double intelligibilité — humaine et technique — ils ont consolidé une ontologie désincarnée de la donnée, maintes fois mise en scène par les commentateurs inspirés de la cybernétique (Blanchette, 2011). Dans le même temps, ils ont invisibilisé le travail qui permet de les produire et de les faire circuler. Ils l'ont relégué au rang de « sale boulot » au sens de Hughes (1958, 1962) : une série de tâches indignes dont on ne veut même pas savoir qui les accomplit ni comment elles sont prises en charge. Comme si le *don* qui était exigé au nom de la démocratie, de la transparence et de l'innovation ne devait pas avoir de coût visible.

Références

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

Baker, K.S. & Millerand, F. 2009. « Infrastructuring Ecology: Challenges in Achieving Data Sharing », *in* Parker E.J., Vermeulen N., & Penders B. (eds) *Collaboration in the New Life Sciences.* Farnham, Ashgate, p. 111-138.

Birchall, C. 2014. « Radical Transparency? », *Cultural Studies ↔ Critical Methodologies*, vol. 14 (1), p. 77-88.

Birchall, C. 2015. « 'Data.gov-in-a-box': Delimiting transparency », *European Journal of Social Theory*, vol. 18 (2), p. 185-202.

Blanchette, J.-F. 2011. « A Material History of Bits », *Journal of the American Society for Information Science and Technology*, vol. 62 (6), p. 1042-1057.

Borgman, C.L. 2015. *Big Data, Little Data, No Data: Scholarship In The Networked World.* Cambridge, MIT Press.

Bowker, G.C. 2000. « Biodiversity Datadiversity », *Social Studies of Science*, vol. 30 (5), p. 643-683.

Carruthers, B.G. & Espeland, W.N. 1991. « Accounting for Rationality: Double-Entry Bookkeeping and the Rhetoric of Economic Rationality », *American Journal of Sociology*, vol. 97 (1), p. 31-31.

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

Codd, E.F. 1970. « A Relational Model of Data for Large Shared Data Banks », *Communications of the ACM*, vol. 13 (6), p. 377-3687.

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

Denis, J. 2011. « Le travail de l'écrit en coulisses de la relation de service », *Activités*, vol. 8 (2), p. 32-52.

Denis, J. & Pontille, D. 2012. « Travailleurs de l'écrit, matières de l'information », *Revue d'anthropologie des connaissances,* vol. 6 (1), p. 1-20.

Denis, J. & Pontille, D. 2014. « Parasite Users? The Volunteer Mapping of Cycling Infrastructures », *in* Mongili A. & Pellegrino G. (eds) *Information Infrastructures: Boundaries, Ecologies, Multiplicity.* Cambridge, Cambridge Scholars Publishing, p. 144-165.

Desrosières, A. 1993. *La politique des grands nombres. Histoire de la raison statistique.* Paris, La Découverte.

Donovan, K.P. 2012. « Seeing like a slum: Towards open, deliberative development », *Georgetown Journal of International Affairs*, vol. 13 (1), p. 97-104.

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

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

Edwards, P., Jackson, S.J., Chalmers, M.K., Bowker, G.C., Borgman, C.L., Ribes, D., Burton, M. & Calvert, S. 2013. *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges.*

Edwards, P. et al. 2011. « Science Friction: Data, Metadata, and Collaboration », *Social Studies of Science*, vol. 41 (5), p. 667-690.

Fraenkel, B. 1994. « Le style abrégé des écrits de travail », *Cahiers du français contemporain* (1), p. 177-194.

Garfinkel, H. & Bittner, E. 1967. « 'Good' organizational reasons for 'bad' clinic records », *in* Garfinkel H. (ed) *Studies in Ethnnomethodology.* Englewood-cliffs, Prentice-Hall, p. 186-207.

Gilbert, G.N. 1976. « The Transformation of Research into Scientific Knowledge Findings », *Social Studies of Science*, vol. 6, p. 281-306.

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

Gurstein, M. 2011. « Open data: Empowering the empowered or effective data use for everyone? », *First Monday*, vol. 16 (2).

Hansen, H.K. & Flyverbom, M. 2014. « The politics of transparency and the calibration of knowledge in the digital age », *Organization.*

Harper, R. 1998. *Inside the IMF. An Ethnography of Documents, Technology and Organisational Action.* San Diego, Academic Press.

Hughes, C. 1958. *Men and Their Work.* Glencoe, The Free Press.

Hughes, E.C. 1962. « Good People and Dirty Work », *Social Problems*, vol. 10 (1), p. 3-11.

Hutchins, E. 1995. *Cognition in the Wild.* Cambridge, MIT Press.

Johns, A. 2009. *Piracy : the intellectual property wars from Gutenberg to Gates.* Chicago, The University of Chicago Press.

Johnson, M.R. 2013. « Material Participation: Technology, the Environment and Everyday Publics », *Information, Communication & Society*, vol. 16 (6), p. 1012-1016.

Kafka, B. 2012. *The Demon of Writing. Powers and Failures of Paperwork.* Brooklyn, Zone Books.

Kelty, C.M. 2008. *Two bits: The cultural significance of free software,* Duke University Press.

Kirsh, D. 1995. « The intelligent use of space », *Artificial intelligence*, vol. 73, p. 31-68.

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*, vol. 40 (3), p. 377-404.

Latour, B. 1993. « Le "pédofil" de Boa-Vista : montage photo-philosophique », *in La clef de Berlin. Petites leçons de sociologie des sciences.* Paris, La Découverte/Poche, p. 171-225.

Latour, B. 2010. « Prendre le pli des techniques », *Réseaux* (163), p. 11-31.

Latour, B. & Woolgar, S. 1979. *Laboratory Life. The Construction of Scientific Facts.* Princeton, Princeton University Press.

Law, J. 1986. « Laboratories and Texts », *in* Callon M. & Rip A. (eds) *Mapping the Dynamics of Science and Technology Sociology of Science in the Real World.* Houndmills, The Macmillan Press, p. 35-50.

Law, J. 2009. « Seeing Like a Survey », *Cultural Sociology*, vol. 3 (2), p. 239-256.

Leonelli, S. 2015. « What Counts as Scientific Data? A Relational Framework », *Philosophy of Science*, vol. 82 (5), p. 810-821.

Lévi-Strauss, C. 1964. *Mythologiques : Le cru et le cuit*, Paris, Plon.

Lynch, M. 1982. « Technical Work and Critical Inquiry: Investigations in a Scientific Laboratory », *Social Studies of Science*, vol. 12 (4), p. 499-533.

Lynch, M. 1985. *Art and Artifact in Laboratory Science. A Study of Shop Work and Shop Talk in a Research Laboratory.* London, Routledge.

Mazzarella, W. 2006. « Internet X-Ray: E-Governance, Transparency, and the Politics of Immediation in India », *Public Culture*, vol. 18 (3), p. 473-505.

McClean 2011. « Not with a Bang but a Whimper. The politics of Accountability and Open Data in the UK. », *Proceedings of the American Political Science Association Annual Meeting*.

Meijer, A. 2009. « Understanding modern transparency », *International Review of Administrative Sciences*, vol. 75 (2), p. 255-269.

Miller, P. & O'Leary, T. 1987. « Accounting and the construction of the governable person », *Accounting, Organizations and Society*, vol. 12 (3), p. 235-265.

Millerand, F. & Baker, K.S. 2010. « Who are the users? Who are the developers? Webs of users and developers in the development process of a technical standard », *Information Systems Journal*, vol. 20 (2), p. 137-161.

Morozov, E. 2014. *To Save Everything Click Here.* New York, Public Affairs.

Myers, G. 1988. « Every picture tells a story: Illustrations in E.O. Wilson's Sociobiology », *Human Studies*, vol. 11, p. 235-269.

Norman, D. 1991. « Cognitive artifacts », *in* Carroll J.M. (ed) *Designing Interaction: Psychology at the Human-Computer Interface.* Cambridge, Cambridge University Press, p. 17-38.

Power, M. 1997. *The Audit Society: Rituals of Verification.* Oxford, Oxford University Press.

Raman, N.V. 2012. « Collecting data in Chennai City and the limits of openness », *The Journal of Community Informatics*, vol. 8 (2).

Ribes, D. & Jackson, S.J. 2013. « Data Bite Man: The Work of Sustaining a Long-Term Study », *in* Gitelman L. (ed) *"Raw Data" is an Oxymoron.* Cambridge, MIT Press, p. 147-166.

Rose, N. 1991. « Governing by Numbers: Figuring out Democracy », *Accountability, Organization and Society*, vol. 16 (7), p. 673-692.

Rosenberg, D. 2013. « Data Before the Fact », *in "Raw Data" is an Oxymoron.* Cambridge, MIT Press, p. 15-40.

Ruppert, E. 2013. « Doing the Transparent State: Open government data as performance indicators », *in* Mugler J. & Park P. (eds) *A World of Indicators: The production of knowledge and justice in an interconnected world.* Cambridge, Cambridge University Press, p. 51-78.

Scott, J.C. 1998. « Seeing Like a State. How Certain Schemes to Improve the Human Condition Have Failed », *in Journal of Social History.* New Haven, Yale University Press.

Shapin, S. & Schaffer, S. 1985. *Leviathan and the Air-Pump. Hobbes, Boyle, and the Experimental Life.* Princeton, Princeton University Press.

Smith, D.E. 1974. « The Social Construction of Documentary Reality », *Sociological Inquiry*, vol. 44 (4), p. 257-268.

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

Strathern, M. (dir.), 2000. *Audit Cultures.* New York, Routledge.

Turner, F. 2006. *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism,* Chicago, University of Chicago Press.

Vollmer, H. 2009. « Management accounting as normal social science », *Accounting, Organizations and Society*, vol. 34 (1), p. 141-150.

Walford, A. 2013. Transforming Data: An Ethnography of Scientific Data from the Brazilian Amazon, Thèse de doctorat, IT University of Copenhagen.

Woolgar, S. & Neyland, D. 2013. *Mundane governance: Ontology and accountability.* Oxford, Oxford University Press.

Wouters, P. & Reddy, C. 2001. « Big science data policies », *in* Schröder P. (ed) *Promise and Practice in Data Sharing.* Amsterdam, NIWI-KNAW, p. 13-40.

Yu, H. & Robinson, D.G. 2012. « The New Ambiguity of "Open Government" », *UCLA Law Review Discourse* (59), p. 178-208.

Zimmerman, A. 2008. « New Knowledge from Old Data: Sharing and Reuse of Ecological Data », *Science, Technology, & Human Values*, vol. 33 (5), p. 631-652.