Rawification and the careful generation of open government data

Jérôme Denis, Samuel Goëta

Rawification and the careful generation of open government data

Jérôme Denis
Centre de sociologie de l'innovation, i3 CNRS - Mines ParisTech,
PSL Research University, France

Samuel Goëta
Département de Sciences économiques et sociales, Telecom ParisTech,
Université Paris-Saclay, France

2017. *Social Studies of Science*, 47(5), p. 604–629

Abstract

Drawing on a two-year ethnographic study within several French administrations involved in open data programs, this article aims to investigate the conditions of the release of government data – the rawness of which open data policies require. This article describes two sets of phenomena. First, far from being taken for granted, open data emerge in administrations through a progressive process that entails uncertain collective inquiries and extraction work. Second, the opening process draws on a series of transformations, as data are modified to satisfy an important criterion of open data policies: the need for both human and technical intelligibility. There are organizational consequences of these two points, which can notably lead to the visibilization or the invisibilization of data labor. Finally, the article invites us to reconsider the apparent contradiction between the process of data release and the existence of raw data. Echoing the vocabulary of one of the interviewees, the multiple operations can be seen as a 'rawification' process by which open government data are carefully generated. Such a notion notably helps to build a relational model of what counts as data and what counts as work.

Keywords

open data, open government, raw data, data labor, administration, invisible work

In 2013, leaders at the G8 summit in Lough Erne, Northern Ireland, signed a charter establishing open government data as the default practice within their administrations. Though the charter represents a rather weak commitment for the signatory countries, it is the temporary result of a gradual process in which government data has become fully-fledged political object. The flow of government data is now considered to be a main motor of transparency, innovation and even democracy. Indeed, open government data is at the center of numerous transparency policies around the world, and the general diffusion of government data has been established as a legal obligation in many countries. Thousands of web portals house vast and varied datasets – disseminated by governments, municipalities, institutions and some corporations – available for public consultation[1]. Data covering extremely diverse subjects, including government spending, the location and nature of bus stops, floods, road traffic accidents, trees' location and healthcare, can be downloaded from there in various file formats.

This hitherto unseen place of data in the public life of contemporary democracies has already received attention. Scholars from several different disciplines have sought to understand the technical, political and cognitive particularities of what some consider a data-driven transparency regime. Insofar as 'content' is concerned, critical studies of the knowledge produced by data, reputed to 'speak for itself', have been numerous. Following Scott (1998), certain scholars have insisted on the fact that transparency based exclusively on quantified phenomena shows a world that is static and largely schematic (Donovan, 2012). Because data are considered to be essentially objective (Birchall, 2014), they naturalize points of view and perform realities that minimize the plurality of information (Johnson, 2013). Some studies, particularly following Porter (1996), often questioned the apparent neutrality of calculating tools used in public and private organizations and, more generally, questioned the realism of accounting theories (Rose, 1991; Vollmer, 2009). For instance, Carruthers and Espeland (1991) on double-entry bookkeeping, or Miller and O'Leary (1987) on standard costing and budgeting theories, have shown that the scientific and technical features of accounting devices are always also political features. Quantified accounts perform a specific kind of reality, in which their conditions of production remain invisible and unquestionable.

From this perspective, open government data initiatives are criticized for their technological determinism. As Yu and Robinson (2012) and Morozov (2014) have shown, the open data movement was defined mostly by technical considerations, largely overlooking its own political dimensions. Most pleas for greater government data diffusion – and most open data policies themselves – are driven by a technological optimism that sees in data transparency a means that will mechanically produce broader and better accountability, and automatically empower citizens as a result. Yet the transparency that is performed through data has important consequences, notably in how privacy is defined and negotiated in modern democracies (Meijer, 2009). The focus on data installs a politics of immediation (Mazzarella, 2006) that dismisses most alternative means of disclosure (Birchall, 2014). This politics empowers, for the most part, the 'already empowered' (Gurstein, 2011; McClean, 2011), and dismisses the tacit knowledge that characterizes some poor groups' claims (Raman, 2012).

These critical studies provide precious resources for questioning the general ambition of open data initiatives and for understanding their social and political consequences. Nevertheless, most of them remain at a level of general discussion and do not examine the concrete conditions surrounding data-driven transparency (Hansen and Flyverbom, 2014); we are largely in the dark

[1] Since the launch of data.gov in 2008 and data.gov.uk in 2010, the model of open government data and data-driven transparency has been exported to other countries in the form of key-in-hand 'packages' (Birchall, 2015). At the writing of this article, according to the last edition of the Open Data Barometer (World Wide Web Foundation), more than 50 countries have implemented an open data policy.

with regard to the situated practices by which data circulate (Mazzarella, 2006). To use Woolgar and Neyland's (2013) terms, this Foucauldian vision of governance, mainly understood through its devices and theories, is missing a complementary look at the mundane governance of data. What does it take for data to be truly 'open'? How do people work behind the scenes of open data online portals? And what happens to the data themselves?

This article begins answering these questions by exploring the conditions under which a series of open data initiatives was carried out in France. We will draw on a two-year ethnographic study within several French administrations involved in open government data projects. During this period we made direct observations, conducted in-depth interviews and gathered numerous documents in two French city halls, two inter-communal structures ('communautés de communes'), one local government ('région'), two transportation carriers, and at Etalab, the open data taskforce of the French government. We also gathered public articles and content from participants in online discussions, different actors who played an important part in the emergence of the open government movement, from the Open Knowledge Foundation to Tim Berners-Lee and others. Even though the situations we observed and the documents we analyzed are sometimes very different, we will not make any comparisons here. Instead, we will foreground the shared issues that the people we observed and interviewed face. More specifically, this material will allow us to investigate an aspect of open government data that has been almost completely neglected until now: the insistence, if not obsession in open data policies, on the *rawness* of data. At the core of the main principles of open government data initiatives is a plea for making public not only data in general, but 'unmodified', 'unadulterated' data. But we do not know how such raw data are actually handled in administrations, or even if they exist as such.

We will first briefly discuss this focus on rawness and its implications, then we will foreground two points that characterize the process leading to the opening of data. We will first show that identifying data is a complex task. From the initial research phase to the extraction of specific files, data identification requires collective explorations, discussions, and a series of operations that turn indistinct pieces of information into *data.* We then show that this opening process also involves transformations that reveal an important aspect of open data policies, the need for a twofold – human and technical – intelligibility. Next, we address the organizational consequences of these operations. Indeed, open data operations affect not only data themselves, but also the ways in which work is distributed inside the administrations carrying these operations out. They make data labor more or less apparent, stabilizing a specific ecology of the visible and invisible (Star and Strauss, 1999). Finally, we return to the consequences that studying data labor may have on the very idea of 'rawness'. Echoing the vocabulary of one of the people we interviewed, we conclude that data rawness on the one hand, and data labor on the other, should not be considered in contradiction. Rather, the multiple operations we discovered can be understood as a 'rawification' process through which open government data are carefully generated.

'We want raw data'

The second principle of the G8 charter identifies the kind of data the signatory governments agree to make publicly available:

> Principle 2: Quality and Quantity
>
> …
>
> We will:
>
> • release high-quality open data that are timely, comprehensive, and accurate. To the extent possible, data will be in their original, unmodified form and at the finest level of granularity available (G8, 2013)

Unmodified data used by administrations have been of public interest only in recent years. The call for the diffusion of raw data was first formulated at the Sebastopol meeting in 2007, which laid down the foundations of open data. Among the principles laid out in the meeting's wake, the second underlines the importance of accessing untreated data ('primary' data). The principle stipulated that 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' (Open Government Data, 2007). This principle was backed by Rufus Pollock, founder of the *Open Knowledge Foundation,* who in a 2007 blog post demanded: 'Give us the data raw, give us the data now!' These words were later coined as a motto by Tim Berners-Lee, the inventor of the World Wide Web, who famously incited his TED talk audience to shout 'We want raw data!'

Such a plea for raw data, now an essential principle of open data policies, draws a very specific picture of government information. Above all, it poses data as a fully-fledged public informational entity. In the words of Berners-Lee, in the text of the G8 charter, or in the principles laid out after Sebastopol, data appear as a new entity emerging in the landscape of public information. The data being requested here are not the archived files whose release to the public has been a cornerstone of modern democracy since the French Revolution (Kafka, 2012), nor are they public statistics whose production and diffusion are already at the heart of contemporary governmentality (Desrosières, 1993). Even though there is no precise definition for the 'data' in open data, the word would seem rather to designate a separate kind of informational artifact, one that, in a way, precedes archives and statistics. Calls to open data do not act only as designators of previously ignored informational objects, but are also founded on the idea that most administrations contain large amounts of this kind of object, raw data that could potentially be communicated to people who might have a use for it. The widespread use of the 'natural resource' metaphor to describe government data bears witness to this idea. Administrations are 'sitting' on a massive reserve of dormant data no less valuable than the 'oil of the 18th century' (Toonders, 2014). Moreover, this insistence on raw data suggests that making it available to the public is a relatively simple process. Given that no changes are permitted (no oil 'refining' is allowed), open data initiatives should involve a simple, almost mechanical operation by which data are made available to the public. This idea of a systematic opening of raw data shares its roots with anti-intellectual property movements, and with cybernetics and the open-source software community (Johns, 2009; Kelty, 2008; Turner, 2006). These roots may be summarized in the slogan, 'information wants to be free.' Not only are government data supposed to be omnipresent in administrations, but they should circulate naturally, free of obstacles and manipulations, so that they may be used in their original state by a wide variety of potential users.

These notions and assumptions recall certain well-known aspects of Science and Technology Studies (STS). Raw data has been dealt with in science for a long time; researchers have pointed to the ambiguity of the term, which appears almost oxymoronic when one considers the daily work necessary for any given dataset to move successfully from one field – or even one laboratory – to the next (Bowker, 2000; Gitelman, 2013). In reality, and despite more and more powerful infrastructures, data never circulate in a perfectly fluid fashion. For example, Edwards and colleagues have pointed to this in their analyses of data sharing and use in large scale, collaborative scientific projects, coining the notion of 'data friction' to describe the difficulties represented by data sharing and flow in scientific disciplines (Edwards, 1999; Edwards, 2010; Edwards et al., 2011). In his examinations of the case of climatology, Edwards uses the term to refer to the costs of reusing data produced in mixed technical and disciplinary settings. To supply models used to measure global warming, for example, scientists must gather records of and collected in a great variety of places and times. Such records have been collected for different reasons, using different methods, and reflect different geographical and institutional circumstances. Thus, gathering the data necessary to calculate the Earth's climate is not a

transparent process that can be reduced to raw information flow. When data circulate, they invariably generate friction.

We are a far cry from an ideal, transparent circulation of data wherein 'raw' information would travel from one department to another, in and out of laboratories, without mediation.

What about government data, though? What kinds of operations are at stake in administrations? Certainly, we can identify similarities with scientific projects studied in information infrastructure studies, but there are also stark differences. It is not possible, for example, to clearly identify the intermediaries of open data circulation: When data are released through an open data program, a wide horizon of potential users emerges, but formal exchanges rarely exist. Another difference lies in the status of the data. In the sciences, the vocabulary used with regards to data is nearly self-evident, in the sense that people working in science are aware that they deal with 'stuff' called data. This is not the case in administrations, where discussion of data tends to be more difficult. Such difficulty translates to a rich terrain for information infrastructure studies. It offers the opportunity to investigate situations in which the circulation of data is problematic, and where the very definition of what data are, look like, or should be, is unstable. In other words, open government data policies are a fertile research area where the ontology of data is explicitly and sometimes difficultly enacted, in the sense of Woolgar and Neyland (2013).

A generative exploration

A first task in the process of opening government data consists in identifying which datasets to open. Early research projects examining the limits of open government data emphasized the delicate nature of data selection (Donovan, 2012; Johnson, 2013; McClean, 2011; Raman, 2012). This research showed that choosing data to open was a politically important process. The decision to open or to keep private certain government data provides the boundaries of transparency, focusing attention on or keeping hidden certain aspects of reality. But this political dimension of data identification, which posits open government data in a binary process between secret and transparency, between closure and disclosure, tends to reduce the identification phase to a selection process. In other words, it assumes the existence of clearly identifiable data and proceeds to examine the circumstances under which these data are sampled. The sampling phase is considered to be the core of the process by which transparency is shaped.

Another kind of identification takes place before, or in lieu of, sampling. Some may see it as too 'technical', all the more so as it holds no clear political weight. However, we believe it is worth understanding this earlier step, in part because it is largely invisible. This invisibility reinforces the apparent 'taken-for-grantedness' of data.

Identification

In administrations, the important thing regarding data, most of or interlocutors told us, is that most people do not know they work with such things as 'data'. In administration, people work with computers, manipulate texts, documents, numbers and tables, but most don't consider that they handle data. This is notable because much of information with which they deal rarely takes the form of crystallized 'products', the shape and movements of which everyone is aware. Therefore, before even considering the data opening process, most of open data program managers need to ask a delicate question: 'What, exactly, are the data we have?' Identification in these cases is not a matter of sampling but of exploration, a crucial and complex operation that

cannot be reduced to a routine gathering of well-defined entities scattered in and around an organization.

In the majority of the cases we studied, the final – sometimes quite distant – goal of identification was to create an inventory that would gather data present in the organization. This inventory would be an exhaustive catalogue, including not only candidates for public release determined on the basis of feasibility and potential interest, but the entirety of the data created by the organization. As an open data project manager explained to us after one whole year of exploring the offices of an inter-communal structure requesting data, such an inventory would greatly help complete data identification once and for all.

> Ideally … we want to have a full list of all the data produced by each department in each local government. … If we could have a full data directory, that would be … it's far-fetched, but that would be great. And then, among the data, we'd identify the sets that could be opened, the sets that couldn't because of restrictions, personal data, sensitive data, etc. (Project manager in an inter-communal structure)

But this vision of an exhaustive identification of data is explicitly seen as utopian. In reality, teams working in administrations never come across ready-made data that can be sorted, counted and classified to produce a general catalogue. Data identification happens slowly throughout the course of exchanges with internal services. The process is progressive, revealing new routes and potential pathways as it goes on. After six months of data identification, this newly appointed open data project manager in a regional government describes the road to data as tortuous. Each meeting with an agent presents new paths to data that he could potentially open:

> We work our way down to the smallest common denominator in order to really identify all the data. What's wild is that after thirty meetings, each time I see them they name five more people that I have to meet with, and, so, all in all, it's growing exponentially. (Open data project manager in a regional government).

The 'gathering' of data consists of an uncertain inquiry during which data identification is progressively completed. Such an inquiry is a collective operation involving interactions with internal services. In the administrations we studied, these discussions and requests varied according to the situation: internal calls, direct requests for communication, open data team suggestions about important data, suggestions made by associations of local citizens or by programmers, etc.

> When we started talking with them about open data, ideas of data sprung up from the services as well as from our side as we had already thought a bit about the question …. For example, when we met the office of services to the local population, we already knew what would be easy for them to open and that it would be interesting to have data regarding names of children born in the city. It is something that had already been done in other cities, so we knew it could be interesting to release the data there, too. We were also interested in data regarding elections, because this is something that the city possesses precisely, such as figures per poll station, etc. And this is something I think was not really done elsewhere. … So it is true that we came to them saying 'this, this is what is truly interesting' but it was overall an exchange. The same with service X, we came asking them, 'what is possible for you to open initially?' We went step by step because it was all quite complex. (Open data project manager in an inter-communal structure)

Data inventory is therefore by no means mechanical. It draws on meetings, discussions and negotiations, which weave together various elements: sharing best practices with other

governments or cities – or, sometimes, standing out from the crowd with new datasets – the significance of certain data, technical difficulties or facilities related to data release and circulation, etc.

This process of progressive identification, and the collective inquiry on which it depends, shows that the identification of data constitutes a technical, organizational and political generative exploration. Data are not immediately workable discoveries made amidst a mass of other available information, and they cannot be selected based on simple criteria defined in advance; rather, information progressively becomes data, and open data, over the course of an internally driven inquiry via negotiations and often-conflicting ideas. The process assembles a perimeter of data that are not only established as open or openable, but also as 'data' in the first place.

Extraction

Identifying the potential data to be opened does not mean that those in charge of an open data program can easily access them. 'Grasping' data is not a simple matter either, mostly because data remain contained in databases. Their 'release' requires them to be extracted from the software and interfaces that used to make them accessible and intelligible to their users. Indeed, relational databases provide dedicated interfaces, called 'user views' (Codd, 1970) that allow a variety of uses (Dagiral and Peerbaye, 2013), while preventing access to the physical organization of the data (Castelle, 2013). The users are dependent on these views, and database software is rarely equipped with an automatic extraction feature that would allow for gathering data independently of the software interface. Thus, after the exploration process dedicated to identifying administrations' data, another exploration begins in order to extract the data from their database. To recover the data 'themselves', data managers must understand the means by which data have been organized and stored. They must go beyond visualization interfaces and down into the guts of hard drives, where the roots of databases lie. This is what a database manager in a transportation carrier explained in the very beginning of his interview: even in a company that explicitly deals with databases and clearly identified datasets, accessing and exporting data take efforts and ingenuity.

> You have to understand that what's complex is that, at the beginning most of the systems and software we bought here, they are absolutely not conceived for doing open data. So it's complicated. We have to come up with these hacks, all sorts of things that allow us to bring out the data properly. (Database manager in a transportation carrier)

Hence, in the vocabulary of relational databases, during the process of extraction, technicians strive to bypass the 'user view' to discover the 'physical view' of the database. Even if broad principles can be found in the way data are actually stored on hard drives, this physical view is always specific and extraction tools generally have to be custom-made. A database manager in an inter-communal structure highlighted this view when we met him, using a domestic metaphor to make us understand what extraction work is about. From one department to another, he encounters very different systems in which data are organized, sometimes in mysterious ways. He has to explore each of them in order to understand their rules and find a way to extract their 'content'. In his terms, even though the organization in which he works could be considered a home, each database resembles a personal closet or drawer he has to examine in order to build a specific tool capable of extracting the data it holds.

> You have to understand that nothing is universal in this stuff. The way you sort your data is like the way you sort your socks at home, everyone can sort them in a different manner. We all have the same drawer but we all sort them differently. (Database manager in an inter-communal structure)

Customization is made more complex by the fact that heterogeneous databases and software often coexist within organizations. As well as multiple versions of the same software. In some cases, IT engineers face a veritable army of data-related devices.

> Another problem … is that because each software is unique, the format and breakdown of the data is all different, so a method you used for one software won't work for another, even if you try and cut down to the basics. The body of it is more or less the same, but the information isn't going to be stored in the same way. So you need to redo an analysis for each new database. And there must be at least fifty different databases. So it's an extremely long process to extract the data. At the town hall, we have data dating back about thirty years, which were stored at the time on big IBM systems which are different from Windows, which are different from Linux. We've got just about everything at the town hall. And so it becomes very complication to extract something specific from them. (Database manager in a town hall)

The homemade explorations and hacks used to extract data provide an idea of the sociotechnical mass out of which the data must be pulled. Furthermore, this battle against software raises questions that are not only technical. Indeed, the explorations used to extract data influence relations between IT departments and their subcontractor service providers.

> The software is made by an American firm that has something like three clients in France and doesn't really care too much about us. The software is completely opaque. What I mean is that our teams don't really know what's happening inside, they don't know what the software does and what it's used to produce. They can't really get inside of it, either. They can't, for example, directly access the database. They have to use a form that the service provider very nicely gives us. We could argue that the data is ours, but because there's this software that's kind of holding them back, they haven't been able to do everything they wanted. For example, the gardens services wanted to put together a system where they could access the contents of the nursery in real time, see which plants were inside, and get all of the information about them, including about their growth. Are the plants ready to be planted, do they need more time? And we realized that it was complicated, because we can't access the database and we can't just go in there and look at what we want. And so, now we're working with them to try and get around this program so that the gardens teams will be able to access the database in certain ways, but it's not easy. (Open data project manager in an inter-communal structure)

According to the terms of the contracts and depending on the good will of the subcontractors, access to data 'themselves' is more or less easy to obtain, and the homespun methods used to get there are more or less akin to misappropriations, or even breaches of contract. This point is essential in interrogating the idea that government data are like sleeping resources, which must only be 'freed' in order to be exploited. This vision of data as a 'commodity' (Ribes and Jackson, 2013) seems to overlook not only the costs involved in extracting it but also the complications brought on by the business relationship between the software providers and the institutions they service. Indeed, these providers own the database paths and the storage systems of their databases: inaccessibility of data is thus at the core of their business model.

For administrations, a large part of extraction work thus consists in regaining control of their data. This is another side of the generative exploration that grounds the opening process. To do this, they disarticulate the sociotechnical agencements that tie data to the private companies that provide the database software. This disarticulation reveals – from a very practical perspective –

the interwovenness of informational infrastructures. As Star and Ruhleder (1996) show, each infrastructure is laid upon another and interacts with others, such that any description claiming to isolate one must be challenged. The opening of government data touches briefly on an isolation of this type, but again, it is fleeting. Data are extracted and isolated from the databases that housed them and, until that moment, had made possible their basic access. When data are opened, though, they are moved and inscribed in a new agencement dedicated to their future release.

Transformations

Though tasks closely linked to identification and extraction are crucial in the gradual generation of government data, they do not take datasets to the point where they can be published as open data. In following the implementation of various open data policies, we observed that data themselves were subject to manipulation. In the process of opening, data may be dramatically transformed. Schematically speaking, there are two categories of manipulations: cleaning of all kinds and interventions that are meant to enhance data intelligibility.

Cleaning

Vocabulary around cleaning is used largely in the sciences, and cleaning operations have been studied in many research projects. For example, many scholars have analyzed the processes leading up to academic publications as an ongoing cleaning operation. Isolated and consolidated, 'clean' scientific results are presented only after the messy conditions of their production have been erased (Gilbert, 1976; Knorr-Cetina, 1981; Latour and Woolgar, 1979; Law, 1986; Lynch, 1982; Myers, 1988). A kind of cleaning is also at stake with data themselves, as researchers try to separate the errors and bias generally caused by instruments from the relevant traces (Latour and Woolgar, 1979; Lynch, 1985). Recently, the growing number of large-scale international and interdisciplinary projects have placed data exchange at the heart of science practices (Borgman, 2012; Leonelli, 2012; Wouters and Reddy, 2001) and bringing data cleaning to a new level of importance (Zimmerman, 2008). For instance, as mentioned already, Edwards has shown that moving from one discipline to another (in his study, from meteorology to climatology) required specific cleaning operations (Edwards, 2010). Walford (2013) explores the complexity of these operations, drawing on an ethnographic study of scientific research carried out in the Amazon.

In the open government data projects we observed, cleaning had several dimensions. First, it was about identifying and correcting mistakes within datasets: values that are considered abnormal and 'holes' in files (blank values). Cleaning also meant harmonizing data; as we have seen, databases, and so datasets, sometimes appear in conflicting formats and versions within the same organization, used by departments that produce and deal with them in locally specific ways. Thus, entities that are otherwise identical can appear in the databases with different units and even different identifiers. As in the case for sharing scientific data on a large scale (Baker and Millerand, 2009), producing open datasets involves bridging these gaps and building coherence between differences and redundancies within multiple datasets. As an open data project manager explained to us, election datasets are, among others, particularly telling examples of such a process. When she opened the files she gathered in order to release ten years of elections data, she faced manifold tiny differences in how the spreadsheets were organized, in how categories and names were used, and so on. If each file were used independently, these differences wouldn't be noticed. Put in a same open dataset, they became serious flaws. Dealing with such mundane

differences and bringing coherence into data from various times and spaces are crucial dimensions of the cleaning processes that open data projects draw on.

> Typically, in the elections dataset, between the files from the previous elections and the old stuff (we went back to 2004), files were not presented the same way. It was very minor stuff but sometimes the column name was the name of the candidate or sometimes the name of the party or both, so I tried to harmonize all that, so that all the files look similar and are structured the same. (An open data project manager in an inter-communal structure)

The resolution to release pristine data, free from mistakes or redundancies, is an important aspect of open data. Indeed, open government data policies literally challenge some data, which, if published as stood, would be perceived as poor in quality even though primary users within the administration have nothing to say against it. This is a widely discussed issue in STS and beyond among scholars who study 'bad records' (Garfinkel and Bittner, 1967) and 'false numbers' (Lampland, 2010). In organizations, data are used daily in very specific settings and for very specific concerns. As 'business data', they are not accurate, true, or high-quality in and of themselves. In their day-to-day uses, what others would call the low degree of their precision, or their lack of harmonization, have no impact on their efficiency. Sometimes, even the opposite is true. There are many 'good organizational reasons', in the words of Garfinkel and Bittner, for this kind of data to persist, and above all to be considered efficient, because accuracy and even 'truth' are always indexical. The data are grounded in the practices of those who manipulate and mobilize them. In the migration of data from one setting to another, issues that were irrelevant in the initial context of use become central. Absences that were never noticed turn into negligence. Approximations or duplicates without importance become mistakes or redundancies. Cleaning and transforming data are thus among the prices to be paid in order to successfully transport data from one setting to another, and in order to avoid perceptions of 'false' or 'bad' data.

The idea of cleaning also emphasizes an essential aspect of open data. Cleaning is a first step in the generation of data that ultimately will be not only unstained, but also generic and universal. These data have to be available for virtually any use. Such a quest for universal usability is not a straightforward journey. It is actually an important feature of a more general requirement, which is at the heart of even more radical transformations: data intelligibility.

The two horizons of data intelligibility

The intelligibility of data is at the core of many texts aimed at guiding open data policies throughout the world. The G8 charter we cited earlier is an example; intelligibility is key to its second principle.

> • ensure that information in the data is written in clear, plain language, so that it can be understood by all …
>
> [Technical annex] Principle 2: Quality and Quantity
>
> 6) We commit to releasing data that are both high in quality as well as high in quantity. When releasing data, we aim to do so in a way that helps people to use and understand them. (G8, 2013)

Though it may seem obvious, perhaps even trivial, this requirement is far from self-evident. In fact, nearly all of the datasets we examined as part of our study failed to respond to this criteria before significant transformations were carried out. The data these sets contained were full of oddities, opaque terms and indecipherable acronyms.

There is a way to improve data intelligibility without directly transforming data: producing metadata such as dictionaries, which facilitate data comprehension. However, as studies in information infrastructure studies have shown, producing metadata is not only expensive but is destined to remain incomplete (Baker and Bowker, 2007; Edwards, 2010). Metadata itself is not easily and universally intelligible, and where 'data frictions' are numerous, 'metadata frictions' are infinite (Edwards et al., 2011). Furthermore, the boundaries between data and metadata are never as clear and as solid as they first appear.

Besides metadata, intelligibility may be attained through direct manipulations and modifications of data. Certain terms are replaced by others, columns are switched around, and information may be deleted. The most common and glaring case is that of acronyms and abbreviations. Professional writing consists largely of shortened forms (Fraenkel, 1994). Organizations are replete with reduced forms of language, often made fun of by outsiders. In the data opening process, everyday acronyms and abbreviations are treated as gaps to be filled in. In transportation, and more generally in every service dealing with locations, maps and street names, this dimension of data opening represents a challenge. As a database manager explains, when it comes to shortened forms and acronyms, the process of opening data goes through an 'unfolding' work, through which words that were only understood by occupational communities become meaningful to everybody else:

> Ninety percent of it is purely technical data, for example, with street names, instead of 'Boulevard du Général de Gaulle', you're going to have 'Bd GDG.' Because the departments who use it to work on the schedules will know what that means. The average passenger has no idea what that means, though, and so we had to cross this data with others that spell out the full street names. (Database manager in a transportation carrier)

As in data cleaning, the main goal here is to make the shift from a specialized setting of meaning where words, acronyms and abbreviations have significance within a largely stabilized language economy, to a broader setting understood by a greater number. As one might suspect, this is no easy feat. Indeed, there is doubt, discussion, and a general sentiment that the process is never complete. All agree that it is impossible to produce data that is universally understood.

But the problem of intelligibility is not resolved through these interventions only. Bringing data to a broader setting is only one of the aspects involved in intelligibility. Open data advocates speak of another crucial dimension, a 'principle' that guides policy and is even used to evaluate it: the need to produce 'machine-readable' data. This is explicit in the G8 charter, the fifth principle of which requires 'machine-readable' data that allow automated processing, or in the principles published by the Sunlight Foundation in 2010, 'information should be stored in widely used file formats that easily lend themselves to machine processing' (Sunlight Foundation, 2010).

Open data thus must not only be intelligible to humans, but to machines as well. This is one of the prerequisites for the free flow of data and one of the central postulates of the plea for the release of raw data. Data is supposed to populate the computers in various departments and should be able to migrate nearly automatically to other computers without being changed. But, as we have seen with regard to extraction, things are not so simple. Data exist in various formats, more or less open, and more or less accessible. In addition, 'machine-readable' is a vague term. How is this requirement enforced? In the practices we observed, technical intelligibility was applied through the adoption of standard or shared formats (Goëta and Davies, 2016). And so it is only once data have been identified, extracted, cleaned and partly transformed that they may truly be considered 'open'. The 'simple' formatting of a dataset is, however, never a straightforward operation. Formatting and reformatting are delicate processes that always have an effect on the data being dealt with.

We will not go into great detail concerning formats. Formatting is a complex subject just now emerging in the field of open data (Goëta and David, 2016). Instead, we propose to focus on two cases in order to emphasize the consequences of even the slightest reformatting process: CSV and GTFS.

In most of the situations we were able to examine, the targeted format was a basic one shared by many actors in the field, rather than a strict, complex format. This basic format, the .csv or 'comma separated values', is considered readable by the greatest number of machines. The production of a .csv file is accessible to almost everyone, and it is possible in many software applications to create a spreadsheet in this format via the 'save as' or 'export' function. Nevertheless, this formatting operation is far from transparent. Carelessness could lead to permanent corruption of a dataset. As its name suggests, the .csv displays values that are separated by commas. No other information can be included. Exporting to this format thus 'flattens' data that, as we have seen, exists on a daily basis in rich and varied forms. Some of the operations we have described are partially aimed at producing a .csv file. Cleaning, for example, aims not only to remove data that might be considered wrong or false for different reasons, but also removes empty fields or hidden rows. This kind of feature may have been intentional, adapted to a specific set of professional activities. Nevertheless, these empty fields and hidden rows are incompatible with .csv format.

Other operations carried out in the name of formatting directly influence the documents used inside departments. Operations such as merging cells or using colors to create reference points in a table must be abandoned. This may have serious consequences. Such functions are invested in the informational and perceptive properties of the screen space. Research on distributed cognition has shown how useful they can be in accomplishing work (Hutchins, 1995; Norman, 1991). Exporting to .csv requires that all spatial encoding of information (Kirsh, 1995) be done away with.

The case of General Transit Specification Format (GTFS) illustrates other transformations. GTFS is a format imposed by Google and certain partner agencies, and has quickly become a *de facto* standard for transport data. GTFS, which is now an open format, is based on a group of several .csv files contained in a compressed .zip file. Each dataset includes at least six fields describing the transport agencies that provide the data, the station stops, lines, projected daily routes for the whole network, and scheduled stop dates and times. The standard specifications define the content of each field (Fig. 1).

Fig. 1: *Information about stops in GTFS files (excerpt)* (Google, 2016)

According to the practices and forms of agencies, adjustments are made in order to make the shift from a given dataset to a series of GTFS data. The format requires each station to have its own identification number. In some administrative services and transport agencies where a station might not have a specific use, an ID number might not have been given. In addition, sometimes ID numbers for the same station differ from one database to the next. During our study, we encountered a transport agency in which stations had different names depending on the department: The network map department used one set of names, where the bus scheduling department used another, yet only the interaction of both databases would provide the information necessary to build a GTFS file. Thus the databases had to be 'corrected' with uniform station names.

The exporting of a dataset to a format as basic as .csv, or to a more complex format such as GTFS reveals a new series of transformations that are added to the other operations we have described. Put together, these processes show there is no mechanical 'opening' of data (as if a kind of tap could be turned on and data would flow out) as much as a delicate manufacture of

Field Name	Required	Details
stop_id	Required	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	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	Contains the name of a stop or station. Please use a name that people will understand in the local and tourist vernacular.
stop_desc	Optional	Contains a description of a stop. Please provide useful, quality information. Do not simply duplicate the name of the stop.
stop_lat	Required	Contains the latitude of a stop or station. The field value must be a valid WGS 84 latitude.
stop_lon	Required	Contains the longitude of a stop or station. The field value must be a valid WGS 84 longitude value from -180 to 180.
zone_id	Optional	Defines the fare zone for a stop ID. Zone IDs are required if you want to provide fare information using fare_rules.txt. If this stop ID represents a station, the zone ID is ignored.

informational artifacts that become open data. Moreover, these transformations rely on invisible work that appears to contradict totally the initial demands for unmodified, primary data. In the next two sections, we propose to discuss these points further. First, we examine some organizational consequences of data labor and its visibility or invisibility. Second, we show how the very idea of 'rawness' can be respecified when data labor is made more visible.

Organizational consequences

The tasks we identified show that data are transformed during the opening process. But, as we will show, administrations also are transformed through these operations. This is true of identification. Creating an inventory is rarely seen as a single, isolated operation, but rather as the first step in a series of data flows, from the services in charge of producing the data and modifying them to those in charge of rendering them accessible to the public. An open data project manager in a local government explained that, in the future, data managers will master the ins and outs of open data so that the methods last and will not require systematic intervention in order to make adjustments:

> We're thinking about this, because for the moment we're dealing with data basically at the end of the line. So the data is picked up somewhere, it gets transferred from service to service, and then we get it. We extract what we need and we publish it on the IT portal in its modified form. And we're starting to

> realize that we're going to need to establish some basic data culture. (An open data project manager in a town hall)

When data flows are established, specific types of data are identified. Also identified, though, are specific locations within the institutions and people in charge of the data and its flow. As with the identification of data itself, these organizational identifications are complex. There is no existing organizational chart, available to all, from which pathways for data flow could be easily determined. Furthermore, open data redistributes certain roles in the organization. New responsibilities are given. The more equipped the inventory is – that is, the more it is inscribed in stable technical infrastructures, even in workflow software – the more it solidifies and reinforces such organizational transformations.

Just as in collaborative scientific projects, improving data intelligibility requires that skills and positions in the division of labor be invented (Baker and Millerand, 2009). The production of metadata, the translation of idiosyncratic names and professional categories into more generic ones, and the harmonization of data identifiers are new tasks that must be introduced to the organization. Room for 'data guys' (Edwards, 2010) has to be made. But these operations and the issues they raise may have even more important consequences for the organization itself. In some administrations, some of the operations we have described coincided with organizational shifts beyond open data teams. Such reorganization is mainly oriented toward the reduction of downstream work on data, through its partial integration upstream. New kinds of collaboration were created, and new steps were created in the initial management of data:

> People don't get that one of the side effects of open data is that in the process, our information system gets more reliable and our data improve in quality. And that's crucial for being able to develop new information systems. [Open data] has a lot of effects we didn't necessarily anticipate at the beginning ….

> Sometimes, it's precisely what we were missing in terms of communication with other services. We realized, for example, when there was a station name change, the person in charge of making the change didn't always make it known. The information wasn't necessarily getting transmitted to the scheduling services. And that's why, sometimes, the station name I had in my file didn't match; it was because I hadn't been told about the change. And so I identified the problems and told a few different people who put together an information process that said, 'when I change a station name, I send a message to so-and-so'. There you go. It's a little thing but in the past it was invisible and it wasn't a big deal. The name of the stop didn't really matter, what mattered was being able to come up with the schedule. (Database manager in a transportation carrier)

Integrating some of the transformations in upstream practices is generally presented as a means to modernize and rationalize administrations. But these reorganizations are not just a product of the new value attributed to data labor, which was previously invisible and is now part of an internal process. These reorganizations transform data labor and reverse the movements described earlier, integrating issues of data opening into activities that previously only used data on an ad hoc basis. Placing formatting or cleaning operations upstream amounts to forcing people to work with generic data considered as 'good data', meanwhile losing the indexical qualities of local and situated data. In other words, in this configuration, open data is not conceived of as the result of dedicated operations, but as a unique horizon common to each activity within a given administration, independent of the specificities of jobs and data, a horizon that should be internalized. This kind of inversion may generate situations like those Garfinkel and Bittner describe in their paper on 'bad records' (Garfinkel and Bittner, 1967), where tensions between barely compatible frameworks of meaning tend to interfere with professional practices.

Therefore, it seems that there are two main ways to take the generation of open data into consideration in these reorganizations. In the first one, data labor is acknowledged as a crucial part of the opening process, the cost of which represents an investment. This implies the creation of new positions and the redefinition of some others within the organization. In the second one, these operations are considered waste and everything is done to minimize their costs. These two directions drive different organizational consequences, but they also bear witness of two ways of understanding what data are. In the first one, the multiplicity of data and their formats is considered inevitable and the coexistence of different versions of data and datasets in the organization is nothing of which to be afraid or ashamed. Conversely, the second case draws on an essentialist definition of data, the generic property of which is seen as an asset. Here, open data programs are seized as occasions to fight idiosyncrasies, to cut the costs of data (re)generation, and to force each part of the organization to work with the same 'open-ready' data.

Rawification

Let us return to data 'themselves.' How should we understand the processes of identification and extraction, and the transformations they undergo if we bring these operations to bear on open data advocates' obsession with 'unmodified' data? What should we do with the idea of 'raw data' given what we have observed?

We may find a clue to answer these questions in the way our interviewees themselves deal with the issue. Though the operations were at times described as difficult, and though their invisibility, non-recognition and lack of means seemed distressing, discussion of them never led to a direct challenge of the notion of raw data. It is tempting to see in this apparent contradiction a conflict between the political and theoretical pretensions of diffusing unmodified data and the description of a series of operations, which aim precisely at modifying data in order to successfully release them. Yet, this would oversimplify what we saw in the previous pages. An open data project manager displayed a very interesting position regarding this seeming conflict when she explained in her own words what it meant to work on data. Instead of opposing her labor to the image of raw data as it should be, she presented it as the very condition under which raw data could exist, using the example of a file she had to modify before it could generate open data:

> This Excel file, they [the agents] worked on it. I have to say that their process is a bit complicated. Basically, their software delivers figures and they annotate it in a file to establish their global statistics. So it really was their working document. However, that is not what we wanted. What we wanted are the rawest data, which is no comments, no charts, no formatting, really the day-to-day data, statistics. This was my job, *I had to re… re-rawify [these data], actually, so the developers could more easily use them*. (Open data project manager in an inter-communal structure)

Thus, if we take what this manager says seriously, opening data implies *rawifying* them. Following this lead, we think, can help challenge how social sciences and notably STS traditionally apprehend raw data.

In the same way that STS has developed a solid critique of the distinction between science and 'bricolage' (Latour, 1993a), the field has also strongly challenged the opposition, borrowed from Lévi-Strauss, between the 'raw' and the 'cooked'. Lévi-Strauss (1983) uses this distinction as a means of separating 'unsocial' raw things from things that become social through cooking. Yet, early laboratory studies and large-scale historical studies have shown that scientific results are not 'harvested' passively, but are rather manufactured through a series of operations that progressively transform data into an academic account (Knorr-Cetina, 1981; Latour and Woolgar, 1979; Shapin and Schaffer, 1985). Above all, there is no 'raw' 'unsocial' data, in the sense of Lévi-

Strauss, before this production of scientific results (Bowker, 2000; Gitelman, 2013). Things are always entangled in the sociomaterial setting that ensures their observability, and cooking data involves dealing with several biases that show that what is treated as 'raw' is always already socialized.

In her study of the Brazilian Amazon, Walford examines the life of what scientists call raw data, which notably involves mistrust with regard to instruments used to 'harvest' data, as they leave traces of their presence that can pollute or stain data. Transforming raw data into data that can be shared involves thus removing the 'artifacts' that affect their meaning, and interfere with access to observed reality. The people who collect raw data, Walford shows, consider them as equivocal entities that need to be disambiguated (Walford, 2013). Hence, in such a configuration, data cleaning amounts to reducing data equivocity.

There is an obvious link between such cleaning processes and open government data programs. Yet, there is an important difference. In sciences, the goal of data cleaning is to make the shift from raw, potentially biased and overly general data to certified data capable of moving from one setting or one discipline to another. Data cleaners narrow the signification of data and reduce 'the extent of background noise … against which an apparently coherent signal can be presented' (Latour and Woolgar, 1979: 37). Rawness here is a problem, and never, in this configuration, are the data manipulated and transformed to *become* raw. On the contrary, the production of scientific results and/or the sharing of datasets requires the technicians and researchers to de-rawify the data, in order to sharpen their meaning and frame their uses.

In open data programs, the process is reversed. Data destined for public use have already had a long social life. Such data do not appear, new and rough, as the simple product of generative instruments, since they have already been in use. These data are already inscribed in sociotechnical agencements that stabilize them and anchor them to specific practices. They are, in a way, already sharpened, and regarding what their opening requires, they are considered too narrow. The tasks that are carried out in view of releasing data to the public thus do not aim to reduce the scope of these data by removing contextual and instrumental residues. On the contrary, the opening process occurs through disembedding data from previous sociotechnical agencements[2] and widening their possible uses. In progressively eliminating what renders data 'business data', data managers strive to transform information into intelligible, though ambiguous, plurivocal data, opened to new, unknown and uncontrolled kinds of treatment. In other terms, if in the sciences, raw data are narrowed and progressively transformed into universal *results*, the generation of open data involves the widening of situated results (working documents, business data…) into supposedly universal *raw data*.

Another stream of research, examining informational infrastructures, has focused on data accumulation and circulation in sciences (Borgman, 2012; Edwards et al., 2011; Leonelli, 2012; Strasser, 2011; Wouters and Reddy, 2001). These studies show that datasets that have been in use must be worked on before they can move from one discipline to another. Such work is crucial to reduce the 'frictions' that data flow inevitably generates (Edwards et al., 2011). Open government data programs are quite comparable to the situations that these scholars have studied. Sharing and disseminating data that have been used for activities dating back, in some cases, decades, requires numerous manipulations and transformations that ensure their intelligibility by future users with specific instruments and in the presence of new data.

---

[2] Yet, of course, it is important to recall that such a disembedding does not amount to freeing data that could exist in and of themselves. The opening process involves a new sociotechnical agencement: as we showed, the generation of open data implies new forms of reduction and closing, specific formatting and other technical standardization procedures.

Such operations occur outside strictly scientific settings. In his ethnography of the International Monetary Fund, Harper (1998) observed a similar type of dynamics in the context of officers' missions. During informal discussions and institutional meetings, IMF officers select and gradually transform data given to them by representatives in the countries they visit. In so doing, they create trustworthy resources used to produce the economic calculations and political reports published by the IMF at the close of each mission. Harper showed that 'moral transformations' of figures are at stake in this process (p. 227). Some of the operations leading up to the opening of data are comparable to these transformations, except that in this case the moral order established is less about the accuracy and fairness of data than it is about data intelligibility and re-usability.

Focusing on data themselves (instead of their progressive mutation into scientific results), Edwards and Harper, among others, foreground the practical and political importance of data labor. This is also what considering rawification helps to do in the case of government data. The notion helps understand that, contrary to what many open data advocates and some scholars may assume, data are not already there, simply buried in the soil of administrations. Data have to be generated, and it takes work. A lot of different pieces of information have to be designated, picked out, isolated, manipulated, adjusted, etc.

Thus, like sharing scientific data, opening government data implies sociotechnical costs. In the case of open data programs, these costs, and the data labor that goes with them, are specific. This is mostly due to the universalism that undergirds the notion of 'opening'. In the sciences, the costs of sharing are essentially dedicated to the passage between social worlds (Bowker and Star, 1999; Edwards et al., 2011). The datasets in these cases are supposed to be used by more or less clearly identified groups. Sharing is built largely around negotiations between spokespersons, and the positioning and adaptation of data to the practices of future users (Millerand and Baker, 2010). In the case of open government data programs, such users are rarely clearly identified. In fact, their non-identification is generally considered a moral guarantee, a condition to avoid favoring certain communities over others or shaping data with regard to specific concerns rather than the general interest. In the data opening process, as in the production of some volunteered geographical information, users may thus be considered as 'parasites' that should not be taken into account too specifically (Denis and Pontille, 2014). Costs of data opening come largely from this situation. It takes an important amount of data labor to generate universally intelligible data without a flesh-and-blood user. This is what *rawification* is all about: trying to (re)produce data that are not only free of any previous uses, but also of explicitly anticipated ones.

Therefore, instead of placing expectations for raw data on one side and the manipulations and transformations we witnessed during our ethnographical study on another, it seems more appropriate to consider how they connect with one another. Identification, extraction, cleaning and the production of both human and technical intelligibility are interventions that do not counter the plea for raw data, but respond to it.

Conclusions: Between *obtenues* and *données*, when are data?

In this article, we have brought to light the conditions by which open government data policies are implemented. In particular, we sought to understand how the principles and guidelines that frame open data policies, which treat data as a natural resource that should flow freely as an 'unmodified' entity, are translated into concrete actions.

First, we showed that the existence of entities conceived of as 'data' is in itself problematic in certain organizations and departments. As such, the initial phases of open data policies resemble

collective inquiries more than harvests or exhaustive inventories. These inquiries end by identifying pieces of information that are designated as data. In addition, some data also require technical modifications in order to be extracted from the original database in which they are 'stuck'. These initial operations show that the preparatory phases and 'backstage' workings of data opening rely on a gradual, uncertain generation of data themselves.

Second, we examined the transformations that data undergo, once they have been identified and extracted, to prepare them for public diffusion. Not only are data cleaned, they are also modified to meet a twofold standard of intelligibility. On the one hand, they are transformed so that they may circulate from a setting of very specialized signification to a setting of near-universal meaning. On the other hand, data are meant to be technically intelligible, that is, 'machine-readable'. Data are therefore modified, adjusted and recalibrated in order to correspond to the norms that ensure such intelligibility.

The operations described have organizational consequences. Transformations in the administrations where data opening takes place have been observed in some cases; in others, organizational transformations were expected to take place. At the heart of these transformations is the visibility or invisibility (Star and Strauss, 1999) of the work required to open data.

Finally, in light of these operations and processes, we returned to the initial subject of the article: the insistence on releasing 'raw data'. Drawing on a term used by one of our interviewees, we showed why it is interesting to consider the work that is performed on the data as different aspects of the same 'rawification' process. Instead of calling for an abandoning of the vocabulary of raw data, or openly criticizing it, we hold that it is appropriate to take rawification seriously, apprehending raw data not only as an oxymoron for social scientists (Bowker, 2000; Gitelman, 2013), but also as a practical oxymoron with which open data managers have to deal every day. From this standpoint, raw data is not an illusion, but a complex and fragile thing to be manufactured.

Beyond the idea of 'rawness', we think that such a pragmatic approach, which takes actors' practices and vocabulary seriously, can help to reconsider the definition of 'data' itself. Indeed, bringing to the surface the invisible work that leads to the 'rawification' of data first invites us to return to Latour's reflection on the French word *données* (data). Instead of speaking about *données* (given), Latour wrote, we should speak about *obtenues* (obtained) (Latour, 1993b: 188). This appears clearly in our investigations: data are outputs. But apprehending data as obtained-data is just the first step in understanding fully the process of opening data. We must not throw the baby of data out with the water of the bath of their generation: We also need to comprehend what given-data encompasses.

In a recent paper, Rosenberg (2013) returns in detail to the links between data in the mathematical sense of the word and what is *donné* or given (that is, both what is already there, and the notion of given as a rhetorical postulate). The term 'data', he explains, was long used to speak of the substance at the base of analysis and calculation, and did not account for its representational qualities. Data, in this sense, are not direct linked to reality.

> The term 'data' itself implied no ontological claim. In mathematics, theology, and every other realm in which the term was used, 'data' was something given by the conventions of argument. Whether these conventions were factual, counterfactual, or arbitrary had no bearing on the status of givens as data. (Rosenberg, 2013: 20)

Studying open data practices from the inside, as we have done, stresses the importance of this aspect in that we have seen how difficult it is to obtain data ready for public

release. Of course, the carefully shaped data that data workers generate in the midst of the opening process may be considered an outcome, that is, *obtenues*. But these obtained-data are meant to transmute into given-data (*données*) for their future users: data that can be taken for granted. The numerous operations we described are the central features of the transactional process through which data are both obtained for some and given for others. Pushing the French word-play one step further, we could thus suggest that open data are akin to gifts (*dons*), carefully crafted informational artifacts offered up to the community.

We should not let our eyes glaze over, though. This gift is offered under very particular circumstances; circumstances which reveal the expectations of the recipients. The principles of open data, the formats used and the standards that progressively take place, along with the international instruments that evaluate open data policies throughout the world, are devices that help to define (though unstably and incompletely) what counts as data. These devices play a central, constraining role in the 'ontological enactments' (Smith, 1974; Woolgar and Neyland, 2013) of data. This is also the case of the countless position papers, articles and blog posts published by critics of released data; such critics are eager to dismiss data formats, quality, or content. For instance, the Open Knowledge Foundation has a project called 'Bad Data' and regularly publishes on its web site detailed criticisms of data sets.

> 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 (Open Knowledge Foundation, 2013)

Statements like this one draw boundaries between good and bad data, sometimes even explicitly designating what are and what are not data. Like international principles and technical standards, they help us understand that data are transactional artifacts that can never be defined as fixed objects, identified by an unchanging set of characteristics. They invite us to expand what Leonelli (2015) calls a 'relational framework' to non-scientific data, whose ability to provide evidence is not essential. To apprehend the transactional process we described in this paper, we propose to borrow Engeström's vocabulary (1990), and ask, '*When* are data?' instead of '*What* are data?'[3] In these terms, our investigation shows that files, documents, numbers, texts and images only become data when they are able – in particular settings, under specific conditions, and meeting more or less negotiated and adjusted criteria – to be taken for granted by those who, after the transaction, become their users.

---

[3] Inspired by Activity Theory, Engeström takes the example of medical records in a chapter entitled 'When is a tool?' to foreground the diversity of the way these records are apprehended in situation, and to describe 'tools' not as a predefined analytical category but as 'transitional, fluid entities' (Engeström, 1990: 189) the stabilization and collective recognition of which is never an easy matter and is always relational.

Since such a relational framework takes into account the enacted ontologies of data, it allows us to understand that the 'same' informational artifact can be considered data or not, depending on the setting and the stakeholders (Borgman, 2015; Leonelli, 2015). For instance, what are documents for some people (Buckland, 1997) are data for others. Above all, highlighting the transactional relation foregrounds the importance of the situated conditions and criteria that establish what counts as data. The capacity to define and impose some of these dimensions is not held by all equally, and considerable power is conferred on those with such capacity. In the case of open government data policies, the early advocates, who came from the open source and free software communities (Kelty, 2008; Turner, 2006), gradually set down the main principles to which official texts now refer (Goëta, 2016). In assuming the genuine existence of raw data in administrations, calling for their fluid circulation, defining constraining criteria for their quality and requiring both human and technical intelligibility, these early advocates consolidated a disembodied ontology of data inspired by cybernetics (Blanchette, 2011). Simultaneously, they rendered the work dedicated to data generation and circulation invisible. They established a moral economy of work in which data labor is relegated to 'dirty work' in the sense of Hughes (1958, 1962): a series of unworthy yet concrete tasks performed by actors kept knowingly unidentified, and who operate largely in the dark. Such a moral economy is certainly not the only possible, and the relational framework we adopt here invites to explore other settings and to discover the diversity of the possible relationships between what counts as data and what counts as work.

Acknowledgments

The authors would like to thank all the people who answered their questions and accepted their presence in this study. They are also grateful to their colleagues at the Center of Sociology of Innovation (Marie Alauzen, Madeleine Akrich, Vincent-Arnaud Chappe, Liliana Doganova, Quentin Dufour, Antoine Hennion, Brice Laurent, Alexandre Mallard, Fabian Muniesa, David Pontille, Vololona Rabeharisoa, Didier Torny, and Alexandre Viole) whose comments on a previous version of this paper were particularly helpful, as well as the reviewers of *Social Studies of Science* for their useful suggestions. Finally, the authors thank Jill McCoy for her always professional and sensitive translation work.

References

Baker KS and Bowker GC (2007) Information ecology: Open system environment for data, memories, and knowing. *Journal of Intelligent Information Systems* 29(1): 127-144.

Baker KS and Millerand F (2009) Infrastructuring ecology: Challenges in achieving data sharing. In: Parker EJ, Vermeulen N, and Penders B (eds) *Collaboration in the New Life Sciences.* London: Ashgate, 111-138.

Birchall C (2014) Radical Transparency? *Cultural Studies ↔ Critical Methodologies* 14(1): 77-88.

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

Blanchette JF (2011) A material history of bits. *Journal of the Amercian Society for Information Science and Technology* 62(6): 1042-1057.

Borgman CL (2012) The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63(6): 1059-1078.

Borgman CL (2015) *Big Data, Little Data, No Data: Scholarship In The Networked World*. Cambridge, MA: MIT Press.

Bowker GC (2000) Biodiversity datadiversity. *Social Studies of Science* 30(5): 643-683.

Bowker GC and Star SL (1999) *Sorting Things Out: Classification and Its Consequences.* Cambridge, MA: MIT Press.

Buckland M (1997) What is a 'document'? *Journal of the American Society for Information Science* 48(9): 804-809.

Carruthers BG and Espeland WN (1991) Accounting for rationality: Double-entry bookkeeping and the rhetoric of economic rationality. *American Journal of Sociology* 97(1): 31-31.

Castelle M (2013) Relational and non-relational models in the entextualization of bureaucracy. *Computational Culture* 3. Available at: http://computationalculture.net/article/relational-and-non-relational-models-in-the-entextualization-of-bureaucracy (accessed 5 May, 2017)

Codd EF (1970) A relational model of data for large shared data banks. *Communications of the ACM* 13(6): 377-3687.

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

Denis J and Pontille D (2014) Parasite users? The volunteer mapping of cycling infrastructures. In: Mongili A and Pellegrino G (eds) *Information Infrastructures: Boundaries, Ecologies, Multiplicity.* Cambridge: Cambridge Scholars Publishing, 144-165.

Desrosières A (1993) *La Politique des Grands Nombres. Histoire de la Raison Statistique.* Paris: La Découverte.

Donovan KP (2012) Seeing like a slum: Towards open, deliberative development. *Georgetown Journal of International Affairs* 13(1): 97-104.

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

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

Edwards P, Mayernik MS, Batcheller AL, Bowker GC, and Borgman CL . (2011) Science friction: Data, metadata, and collaboration. *Social Studies of Science* 4(5): 667-690.

Engeström Y (1990) *Learning, Working and Imagining.* Helsinki: Orienta-Konsutit Og.

Fraenkel B (1994) Le style abrégé des écrits de travail. *Cahiers du Français Contemporain* 1: 177-194.

G8 (2013) G8 open data charter and technical annex. Available at: https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex (accessed 3 May, 2017)

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

Gilbert GN (1976) The transformation of research into scientific knowledge findings. *Social Studies of Science* 6(3-4): 281-306.

Gitelman L (ed) (2013) *'Raw Data' is an Oxymoron.* Cambridge: MIT Press.

Goëta S (2016) *Instaurer les données, instaurer les publics. Une enquête sociologique en coulisses de l'open data*. PhD Thesis, Telecom Paristech, France.

Goëta S and Davies T (2016) The daily shaping of state transparency: Standards, machine-readability and the configuration of open government data policies. *Science & Technology Studies* 29(4): 10-30.

Google (2016) Google transit APIs > static transit. Available at : https://developers.google.com/transit/gtfs/reference/stops-file (accessed 3 May, 2017)

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

Hansen HK and Flyverbom M (2014) The politics of transparency and the calibration of knowledge in the digital age. *Organization* 22(6): 872-889.

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

Hughes EC (1958) *Men and Their Work.* Glencoe: The Free Press.

Hughes EC (1962) Good people and dirty work. *Social Problems* 10(1): 3-11.

Hutchins E (1995) *Cognition in the Wild.* Cambridge, MA: MIT Press.

Johns A (2009) *Piracy: The Intellectual Property Wars from Gutenberg to Gates.* Chicago: The University of Chicago Press.

Johnson MR (2013) Material participation: Technology, the environment and everyday publics. *Information, Communication and Society* 16(6): 1012-1016.

Kelty CM (2008) *Two Bits: The Cultural Significance of Free Software*. Durham, NC: Duke University Press.

Kafka B (2012) *The Demon of Writing: Powers and Failures of Paperwork.* New York: Zone Books.

Kirsh D (1995) The intelligent use of space. *Artificial Intelligence* 73: 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* 40(3): 377-404.

Latour B (1993a) *We Have Never Been Modern.* Cambridge, MA: Harvard University Press.

Latour B (1993b) Le 'pédofil' de Boa-Vista : Montage photo-philosophique. In: Latour B *La Clef de Berlin et Autres Leçons d'un Amateur des Sciences.* Paris: La Découverte, 171-225.

Latour B and Woolgar S (1979) *Laboratory Life: The Construction of Scientific Facts.* Beverly Hills: Sage.

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

Leonelli S (2012) When humans are the exception: Cross-species databases at the interface of biological and clinical research. *Social Studies of Science* 42(2): 214-236.

Leonelli S (2015) What counts as scientific data? A relational framework. *Philosophy of Science* 82(5): 810-821.

Lévi-Strauss C (1983) *The Raw and the Cooked.* Chicago: University of Chicago Press.

Lynch M (1982) Technical work and critical inquiry: Investigations in a scientific laboratory. *Social Studies of Science* 12(4): 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* 18(3): 473-505.

McClean T (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*. Seattle, USA.

Meijer A (2009) Understanding modern transparency. *International Review of Administrative Sciences* 75(2): 255-269.

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

Millerand F and 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* 20(2): 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* 11: 235-269.

Norman D (1991) Cognitive artifacts. In: Carroll JM (ed) *Designing Interaction: Psychology at the Human-Computer Interface.* Cambridge: Cambridge University Press, 17-38.

Open Government Data (2007) The annotated 8 principles of open government data. Available at https://opengovdata.org/ (accessed 5 May, 2017).

Open Knowledge Foundation (2013) Bad data. Available at: http://okfnlabs.org/bad-data/ex/tfl-passenger-numbers (accessed 5 May, 2017).

Porter TM (1996) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life.* Princeton: Princeton University Press.

Raman NV (2012) Collecting data in Chennai City and the limits of openness. *The Journal of Community Informatics* 8(2). Available at: http://ci-journal.net/index.php/ciej/article/view/877/908 (accessed 5 May, 2017).

Ribes D and Jackson SJ (2013) Data bite man: The work of sustaining a long-term study. In: Gitelman L (ed) *'Raw Data' is an Oxymoron.* Cambridge, MA: MIT Press, 147-166.

Rose N (1991) Governing by numbers: Figuring out democracy. *Accountability, Organization and Society* 16(7): 673-692.

Rosenberg D (2013) Data before the fact. In: Gitelman L (ed) *'Raw Data' is an Oxymoron.* Cambridge, MA: MIT Press, 15-40.

Scott JC (1999) *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press.

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

Smith D (1974) The social construction of documentary reality. *Sociological Inquiry* 44(4): 257-268.

Star SL and Ruhleder K (1996) Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 7(1): 111-134.

Star SL and Strauss A (1999) Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer Supported Cooperative Work (CSCW)* 8(1): 9-30.

Strasser BJ (2011) The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine. *Isis* 102(1): 60-96.

Suchman L (1996) Supporting articulation work. In: Kling R (ed) *Computerization and Controversy: Value Conflicts and Social Choices.* San Diego: Morgan Kaufmann, 407-424.

Sunlight Foundation (2010) Ten principles for opening up government information. Available at: https://sunlightfoundation.com/policy/documents/ten-open-data-principles/ (accessed 3 May, 2017).

Toonders J (2014) Data is the new oil of the digital economy. *Wired*. Available at: http://www.wired.com/insights/2014/07/data-new-oil-digital-economy/ (accessed 3 May, 2017).

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* 34(1): 141-150.

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

Woolgar S and Neyland D (2013) *Mundane Governance: Ontology and Accountability.* Oxford: Oxford University Press.

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

Yu H and Robinson DG (2012) The new ambiguity of ‘open government’. *UCLA Law Review Discourse* 59: 178-208.

Zimmerman A (2008) New knowledge from old data: Sharing and reuse of ecological data. *Science, Technology, & Human Values* 33(5): 631-652.