Exploration, Extraction and 'Rawification'. The Shaping of Transparency in the Back Rooms of Open Data

Jérôme Denis, Samuel Goëta

Exploration, Extraction and 'Rawification'

The Shaping of Transparency in the Back Rooms of Open Data

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

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

'After The Reveal. Open Questions on Closed Systems'
Neil Postman Graduate Conference, New York, February 28th 2014

Abstract

With the advent of open data initiatives, raw data has been staged as a crucial element of government transparency. If the consequences of such data-driven transparency have already been discussed, we still don't know much about its back rooms. What does it mean for an administration to open its data? Following information infrastructure studies, this communication aims to question the modes of existence of raw data in administrations. Drawing on an ethnography of open government data projects in several French administrations, it shows that data are not ready-at-hand resources. Indeed, three kinds of operations are conducted that progressively instantiate open data. The first one is *exploration*. Where are, and what are, the data within the institution are tough questions, the response to which entails organizational and technical inquiries. The second one is *extraction*. Data are encapsulated in databases and its release implies a sometimes complex disarticulation process. The third kind of operations is 'rawification'. It consists in a series of tasks that transforms what used to be indexical professional data into raw data. To become opened, data are (re)formatted, cleaned, ungrounded. Though largely invisible, these operations foreground specific 'frictions' that emerge during the sociotechnical shaping of transparency, even before data publication and reuses.

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

Transparency in action and the production of open government data

For some years now, public administrations "raw data" have entered politics. As it is presented as an essential means for a renewed transparency, an instrument for a new institutional accountability, its diffusion to the widest audience has become a crucial stake for policy makers around the world. Open government data indeed has been at the center of several transparency policies and has been established as a legal obligation in many countries[1]. This led to the creation of many web portals in which one can find extremely various datasets, disseminated by governments, municipalities, institutions and some corporations.

Several studies analyzed open government data policies as a new step towards the generalized transparency that emerges in many fields, feeding an "audit culture" (Power, 1997; Strathern, 2000). In this perspective, the opening of public data and the creation of devices that encourage its reuse both the State, the shape of which becomes unstable, and the citizens, who are positioned as "data publics" (Ruppert, 2013). Because it extends the field of accountability, open data can be considered as a flagship tool of a new form of governmentality, which shapes the figure of an informed citizen (Barry, 2001).

Yet, how open government data actually circulate, and moreover how the data are concretely "opened" remain largely overlooked. The installation of open government data portals, which provide real infrastructures dedicated to the dissemination of data, required important technical and organizational transformations. Within administrations, teams have been created, departments redesigned. What is the nature of these transformations that occurred in the back rooms of open data? What does it takes for the data to actually be "opened"? Who are the invisible workers that handle them and take care of them (Denis, Pontille, 2012)?

To answer these questions, it is useful to step back to what happened in some sciences, in which publishing and exchanging data have become crucial matters. Indeed, data collection and diffusion have been progressively decoupled from the publication of results in many scientific disciplines, becoming a kind of scientific production of its own (Bowker, 2000). This new emphasis on data was accompanied by major shifts in the organization of scientific work (Wouters et al., 2001; Hine, 2006; Edwards, 2010). Following the first Information Infrastructures studies, we think it is useful to bring to the light such concrete issues in the production of open data, insisting notably on the frictions which occur when producing and

[1] The claim to diffuse raw data was formulated at the Sebastopol meeting in 2007, which laid down the foundations of open data. Among the principles, the second underlines the importance to access data without treatment called "primary". « 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. » ("8 Principles of Open Government Data". http://www.opengovdata.org/home/8principles). The Lough Erne G8 held in 2013 led to the signature of a charter, which stipulates that open data will be the default practice of administrations, an obligation combined with detailed action plans in signatory countries.

exchanging data (Edwards, 2010; Edwards et al., 2011), the political aspects of databases management (Bowker & Star 1999) and the necessity to respecify the vocabulary of “raw data” (Gitelman, 2013).

This is a way to understand how transparency is actually shaped in public administrations, even before its diffusion and its use.

The aim of this communication is thus to bring to light the sociotechnical “thickness” of open government data, and to understand the political, technical and organizational issues carried by the transparency it performs. For that purpose, we will build on an ethnographic enquiry made of direct observations, in-depth interviews and documents analysis in several French administrations involved in open government data projects.

Here, we present three crucial steps that punctuate work in the back rooms of open data: identification, extraction and ‘rawification’. These three steps show that data is not an available commodity within administrations, which would be naturally ready for its public diffusion, mechanically producing transparency. The data very existence is far from being obvious, and instead of being an intrinsic property, “rawness” is the outcome of numerous operations.

Exploration

Unlike scientific practices, which are entirely turned into the production of data, administrations rarely consider data as a self-evident matter. Not only the perimeter of the data that are potentially candidates for opening is debated and shifts for each open government data project, but the very knowledge about their existence, their nature and their location is equally problematic. Before asking the question “how are we going to open this dataset or this one?”, the persons in charge of implementing an open government data program are facing an even more abysmal interrogation “what data do we possess?”

In the process of opening public data, identification is thus a crucial and complex operation, which rarely summarizes to the more or less fastidious gathering of well-defined entities that one should simply chase within the organization. Even if some people dream of an exhaustive inventory of all the data that are present in an administration, the gathering actually occurs as an uncertain exploration during which data identification is progressively settled, through interactions with internal services.

> When we started talking with them about open data, ideas of data sprung out from services as well as from our side as we 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 them too. We also were interested in data regarding elections, because this is something that the city possesses precisely, such as figures per poll stations, 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 everything was not simple to do. (An open data project manager in an intercommunality)

Thus, data inventory is by no means mechanical. It implies meetings and discussions, sometimes negotiations, which are linked to various matters. Among these, the degree of sensitivity of data, that is the risk its opening represents, is often crucial. The concern for transparency might encounter detrimental consequences led by the publication of certain data. It is particularly the case of data concerning infrastructures as explained by a civil servant in charge of roads in a French city:

> We know there are some folks with disreputable jobs who will shut down public lighting on some street lamps by fumbling into the cables and will get rid of the right cables so that they can do their business quietly, without having too much light to disrupt them. So if we were to open [data on] cabling, they could know where the juice for each electrical cabinet would come from and where it lights, they would break this electrical cabinet and have a whole neighborhood in the dark to do business quietly. Well, it's a bit far-fetched but well. The risk exists and well, there is not really much interest to liberate this data. (A database manager on roads)

Thus, neither data are obviously available in the services, nor is it selected on simple criteria, which would be defined in advance. Through its progressive and collective exploration, data is co-elaborated. Identification itself is thus a crucial part of the opening process.

Two aspects are important to understand what is at stake with this exploration step. First, it has organizational consequences. These are not only specific types of data that are identified. Places within the organization are also designated in this process, as well as persons, who are set up as responsible of these data and their circulation. Second, as a process, data identification is more a matter of instantiation than it is of genuine designation. This identification process is generative. It engenders a certain reality (Law, 2009), a perimeter of data that is not only established as open ("openable", in the first place) but also as "data" at all.

Extraction

Once the data candidate to opening are identified, one still have to "grasp" them, which is, again, not self-evident. The data do not become available by their mere identification: they remain encapsulated in databases and their release requires them to be extracted from the software that makes them visible to their users.

Relational databases provide dedicated professional interfaces, called "user views" (Codd, 1970). Aimed at simplifying the use of the database, these multiple views allow a variety of uses (Dagiral et Peerbaye 2013), while preventing from accessing to the physical organization of the data (Castelle 2013). The users are thus dependant on these views and rare are the databases equipped with an automatic extraction feature, which would allow gathering the data independently of the software interface. To allow this extraction, people must delve beyond visualization interfaces, in the guts of hard drives, at the roots of databases.

> In fact, you have to understand that what is complex is that, at the beginning most of the systems and softwares we bought here, they are absolutely not conceived for doing open data. So it is complicated. We must, ourselves, develop small tools [*moulinettes*], all sorts of things that allow us to bring out data properly. (A transport database manager)

Furthermore, even if broad principles can be found in the way data is actually stocked on hard drives, this "physical view" is always specific. Extraction tools have to be custom made for each database software, and sometimes each version of each software.

> What you have to tell yourself is that nothing is universal in that stuff. That is 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).

Therefore, to implement an open government data policy, one must be capable, once data identified, to bypass the databases one way or another and to harvest the data directly from their storage space. Such extraction work performs the second step of the progressive instantiation of data. It shows again that raw data is not naturally available in administrations, and that transparency does not rhyme with immediacy.

This point is essential to question the idea that public data would be sleeping resources, which only has to be "freed" in order to be exploited. This vision of data as a "commodity" (Ribes et Jackson 2013) is jeopardized not only by the cost it represents for extracting but also by the ambiguity of what the technical providers deliver to institutions by providing them their databases software. These providers own the database paths and the storage systems of their databases: inaccessibility of data is at the core of their business model. A large part of extraction work therefore consists for institutions to regain control of data by disarticulating the sociotechnical assemblages that tie data to some private companies.

"Rawification"

We know that the term "raw data" is ambiguous and that, once we look closely at the practices it is associated to, it can be likened to an oxymoron (Bowker 2000; Gitelman 2013). It is precisely the case of open government data projects which analysis makes visible the series of transformations data are subject to so that they *become* raw. We will comment here three operations which take part to what some open data people call "rawification" work: reformating, cleaning and ungrounding.

Extracting databases does rarely suffice to obtain data that are ready to be opened. Their opening must come from formatting, which ensures they are readable by the most common tools used by developers, but also by the general public. Obviously, this very idea of a 'suitable' format hides a complexity we will not unfold here. During our enquiry, CSV (comma-separated values) was the most frequent standard, chosen notably because it is an open format allowing its use by every spreadsheet programs. We heard during a meeting starting an open government data project this phrase, which epitomizes how important is formatting in the process of producing open data: "to me raw is CSV." However, translating a dataset into CSV is nothing but self-evident. Each export from a proprietary format from original databases or a spreadsheet software to its desired format leads to a series of problems and requires adjustments which ensure that initial data are not corrupted in the process.

Before or after being formatted, data are subject to another type of treatment: they are "cleaned"[2]. In open government data programs, cleaning concerns several aspects. It first consists in correcting mistakes within the datasets: values that are considered as abnormal, and "holes" in files (blank values). Cleaning also means harmonizing data. As we have seen, databases, and so datasets, appear in different formats and versions within the same institution. They are generally manipulated by departments that produce and deal with data in their specific way. Thus entities *a priori* identical can appear in the databases with different units, even different identifiers. As in the case for sharing scientific data on a large scale (Baker, Millerand, 2009), producing open datasets involves bridging these gaps and building coherence between differences and redundancies within multiple datasets.

> Typically, on the elections dataset: between the files from the previous elections and the old stuff (we went back until 2004), files were not presented the same way. It was very stupid things but sometimes the column name was ever the name of the candidate or sometimes the name of the

[2] The vocabulary of cleaning, associated with the idea of quality data, is widespread in scientific fields. It was recently subjected to an in-depth ethnographic work on the case of scientific work in the Amazon rainforest (Walford 2013)

> party or both, so I try to standardize all of this so that all files look similar and are structured the same. (An open data project manager in an inter-communal structure)

This is an important aspect of open data. Open government data programs literally challenge some data, which if it was published as it was, would be perceived of poor quality, even though their users within the organization have nothing to say against it. This is a widely discussed issue in STS and beyond, about "bad records" (Garfinkel, 1967) and "false numbers" (Lampland, 2010). In organizations, professional data is not accurate or true in itself. The low degree of its precision or its lack of harmonization have no impact on its efficiency, sometimes the opposite. There are many "good organizational reasons", in the words of Garfinkel, that this kinds of data persist, simply because their accuracy and even their "truth" are grounded in the practices of those who manipulate and mobilize them. It is the confrontation between practices with specific issues that might lead to stigmatize such or such data as false or bad. In that sense, open government data programs can be considered as tests. By migrating data to a new framework, they potentially foreground issues that were not much relevant in the initial context of its use. Absences that were never noticed become negligence, approximations or duplicates without importance become mistakes or redundancies. Cleaning data is the cost to pay to avoid these issues and to transport data from one framework to another.

The idea of cleaning emphasizes another essential aspect of the framework towards which data is meant to migrate. As it is cleaned, data becomes generic, cleared from the scoria resulting of the situated activities it used to feed. Cleaning is a first step towards the universality which open data promotes. The deletion of ambiguities and blanks produce datasets suitable to virtually any usage.

Such a projection of data towards universal uses is fed by another operation, very close from cleaning: ungrounding. Besides producing data for which nothing can be reproached in terms of coherence and completeness, rawification work involves erasing the traces of previous uses.

> This Excel file, they 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. (An open data project manager in an inter-communal structure)

Ungrounded, data is delivered from the marks of the practices it initially was tied to. This operation constitutes another dimension in data instantiation. Once more, the scope and the very definition of data is specified while components such as colors, comments and sometimes sub-categories are sidelined, being performed as *non-data*. Thus, following the cleaning of data, ungrounding practices achieve a process of purification by which data progressively becomes "raw". It enacts an essential dimension of the fabric of transparency: data intelligibility.

Therefore, to become open, data has to be raw, but to be raw, most of the data have to be "rawified", that is formatted, clean and ungrounded. These qualities are not intrinsic: they imply complex sociotechnical work which itself requires, as in the scientific collaboration projects, that skills and positions in the division of labor are being invented (Baker, Millerand 2009).

Such operations may have even more important consequences on the organization itself. In some administrations, the steps that we just described are indeed sometimes the occasion to rethink some of the organizational processes, in order to reduce downstream work on data. Integrating some of "rawification" work is generally presented as a means to modernize and rationalize administrations, which is not without raising crucial issues. Placing formatting, cleaning and ungrounding operations upstream is a way to force people to work with generic data, losing the quality of their grounding. In other words, in this configuration, transparency is not thought as the result of dedicated operations, but as a unique horizon shared by every activities in administrations, independently from the specificities of jobs and data. This inversion assumes the risk of installing the same situations Garfinkel once described (1967), where tensions between hardly compatible frameworks of meaning tend to parasite professional practices.

Conclusion

In this communication, we showed that, conversely to what certain injunctions or legal dispositions imply, the government data are not “already there” entities, which would be predisposed to mechanic openness. The ethnography of concrete activities on which open government data programs rely shows that data is progressively instantiated as open data. We have insisted on three crucial steps of this instantiation — exploration, extraction and “rawification”, which underline the sociotechnical operations data is subject to. In sciences, these transformations have been studied as means of mutating *raw data* into refined *certified data*, ready to be handled in specialized activities (Edwards 1999; Walford 2013). In the case of open data, the process is somehow reversed: it involves transforming sets of *professional data*, which already had a long social life, into *raw data*, open to many uses. Therefore, raw data is not a starting point here, but the result of several “delocalization” operations.

To conclude, we may get back to the vocabulary of raw data. Once we highlighted the operations that take part in the shaping of such raw data, should we consider that it does not exist “in reality” and that the mere notion of raw data is a fiction, even an illusion? Undoubtedly not, unless we seek to impose a definition of raw data that is eventually not shared with the persons concerned. Instead, we think raw data should be understood as an oxymoron (Bowker 2000; Gitelman 2013)… but an oxymoron that the data workers who shape transparency on daily basis have to deal with. This is a way to surface invisible work involved in raw data and to highlight the tensions both data and their workers encounter during the series of transformations that lead to raw data, the cost of which is never completely measured in advance.

References

Baker, K S and Millerand, F, 2009, “Infrastructuring Ecology: Challenges in Achieving Data Sharing”, in *Collaboration in the New Life Sciences* Eds E J Parker, N Vermeulen, and B Penders (Ashgate), pp 111–138.

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

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

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

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

Codd, E F, 1970, “A Relational Model of Data for Large Shared Data Banks” *Communications of the ACM* 13(6) 377–3687.

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

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

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

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

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

Gitelman, L ed, 2013 *“Raw Data” is an Oxymoron* (MIT Press, Cambridge).

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

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

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

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

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

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

Ruppert, E, Law, J, and Savage, M, 2013, “Reassembling Social Science Methods: The Challenge of Digital Devices” *Theory, Culture & Society* 30(4) 22–46.

Strathern, M, 2000, “The Tyranny of Transparency” *British Educational Research Journal* 26(3) 309–321.

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

Wouters, P and Reddy, C, 2001, “Big science data policies”, in *Promise and Practice in Data Sharing* Eds P Wouters and P Schröder (NIWI-KNAW, Amsterdam), pp 13–40.