Data and its invisible work

Jérôme Denis

▶ **To cite this version:**

Jérôme Denis. Data and its invisible work. Science + technology by other means - 4S/EASST Conference, Aug 2016, Barcelona, Spain. halshs-01364311v2

HAL Id: halshs-01364311
https://shs.hal.science/halshs-01364311v2

Submitted on 7 Oct 2016

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

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

data and its invisible work

Jérôme Denis
Centre de sociologie de l'innovation
i3 (CNRS 9217) - Mines Paristech
jerome.denis@mines.paristech.fr

4S/EASTT Barcelona - September 2016 (Track *The Lives and Deaths of Data*)

Abstract

In this communication I question the contrast between the image of data as primary material that can seamlessly flow from one place to another and the work that such circulation takes behind-the-scenes, the frictions the data engender, their fragility, and their cost. Of course, following Leigh Star (1999) and Lucy Suchman (1995), one can discuss this paradox by surfacing invisible work. But, drawing on two ethnographic studies (in the back office of a bank, and a start-up that works with French administrations), I argue that such paradox also invites to adopt what Leonelli (2015) calls a "relational framework" regarding data, and then investigate the relationships between what counts as data and what counts as work.

In this communication I question the contrast between the image of data as a primary material that can seamlessly flow from one place to another (and that represents huge opportunities or huge risks, depending on what side you're standing), and the work that such circulation takes behind-the-scenes, the frictions the data engender, the fragility of data themselves, and the costs of their circulation. Surfacing invisible work, and following the path of Leigh Star (1999) and Lucy Suchman (1995) among others, is of course a way to investigate this secret life of data. But I would like to stress that this paradox can also teach us something about data themselves and their situated ontologies, that is the negotiations through which some things can be labelled as data or not.

First I will briefly illustrate the paradox itself with two case studies that examines the place of data and data labor in administrations. I will describe how certain properties of data are invested in, and then bring to the light the backstage work and the mundane life of data.

Then, I will foreground the interest of questioning the relation between data and their work in terms of what Star and Strauss (1999) called the ecology of visible and invisible, and to unfold the links between the situated definitions of what is and what is not data, and the definitions of what counts as work.

standardized information and the emergence of data in organizations

First, let us remind that, from what Jack Goody (1986) termed the graphic reason, to the managerial revolution that occured at the beginning of the 20th century (Beniger, 1986; Chandler, 2003; Agar, 2003) and the mechanization of information work it lead to (Yates, 1999; Gardey, 2008), the emergence of data within organizations is a long story. Today, data seems to be everywhere, staged as a natural resource, a fluid entity that circulates seamlessly through the world.

in this schematic history, even if a lot of things dramatically changed, what is data remained basically the same: stable pieces of information that flow within companies and administration. Particularly important is the role of standardization of these pieces of information. Standardization was progressively invested in to ensure security, efficiency, and productivity (Beniger, 1986; Agar, 2003). Basically, today, data in organizations take the form of what Latour (1987) calls *immutable mobiles*.

mess and work behind the scenes

Yet, in STS and Accounting studies, or Media studies more recently, many scholars have shown that such an account gives a very partial image of the story. As soon as we go backstage, we face workers, and the messy side of standardized information the production and the circulation of which is never an easy matter (Denis & Pontille, 2012).

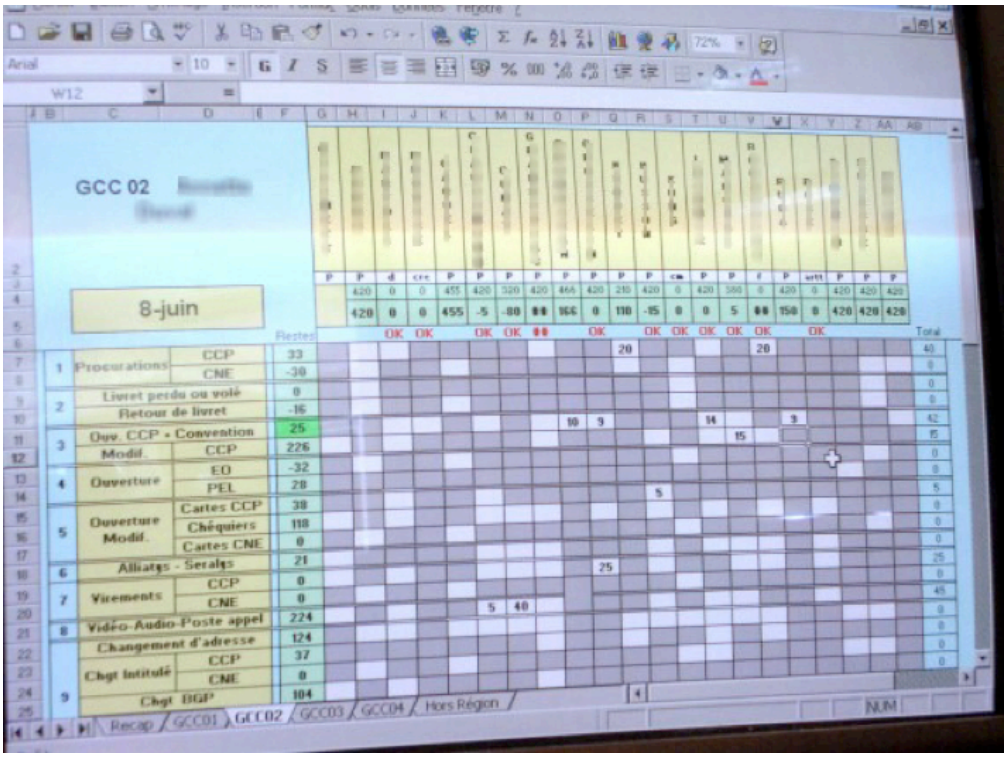
> No matter how good the scheme, its scope is limited by the fact that data entry is never an easy task, and there are never enough resources or trained personnel to make it happen. (Bowker & Star, 1999, p. 107)

I would like to quickly illustrate the two sides of this contrasted life of data with two studies I conducted these last years.

the investment in standardized all-written communication

Let me invite you first in the back-office of a Bank where people had to verify the customers files that arrived from the commercial agencies. This was only once such verification was done that the bank accounts could actually be opened.

Just before I began my observations, the department had been re-organized to fit with the idea that the circulation and verification of files should be fully transparent operations.



The first feature of this re-organization concerned the process of verification itself. Employees were asked to stick to a binary rule. They should simply check if every pieces of information and every documents complied with the standards, and proceed to the data entry of “good” files and send the faulty files back to the agency where it came from.

Il addition, the files had been classified into categories, and for each category, standardized amount of time for verification had been set. These standardized amounts of time notably allowed work to be distributed every morning by the manager through a spreadsheet that calculated the number of files she could give to each of her employees.

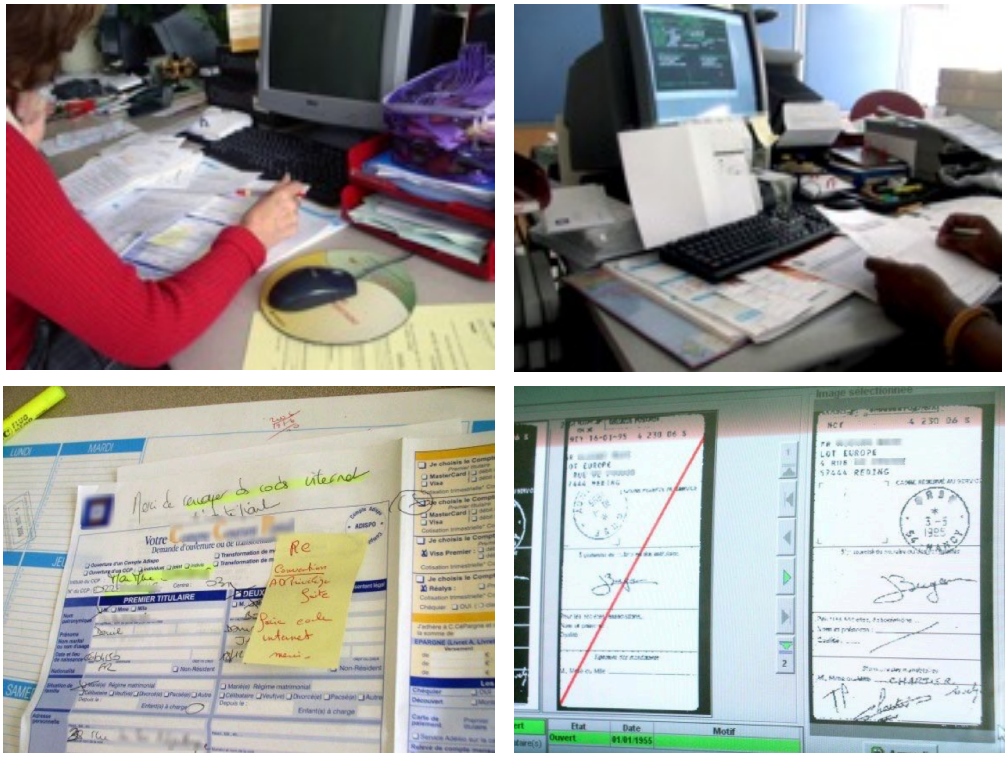
And last but not least, all the telephones of the platform had been removed, and every oral communication with the agencies had been forbidden. Needless to say, before the re-organization, the phone calls would take a great part in the day to day verification work. Calling the agencies was notably and efficient means for employees to clarify the ambiguities of the files and the uncertainties they raised.

Such an re-organization is interesting for it clearly bear witness to a real *data positivism*: an investment in data as standardized and transparent information. Data are in the files and should flow by themselves from the agency to the databases, almost automatically, human workers being there just to give a mechanical glimpse to them.

what verification takes

Of course, the first minute one entered the platform, one would understand that things were way more complicated than that.

Verification would almost never take the form of a binary separation between faulty files and others. In fact, employees regularly engaged into real enquiries through ambiguous data, unclear and even missing information. They annotated the files, browsed heterogenous databases, searched for external information, and juggled with distant authorities.



In short they did not sort data out, but progressively instantiated heterogenous information into something that could be considered as data and was eventually entered into a database.

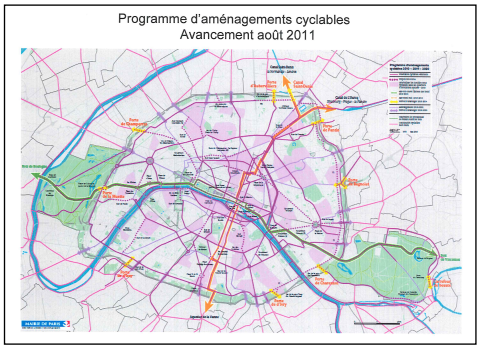
data, already there, naturally available?

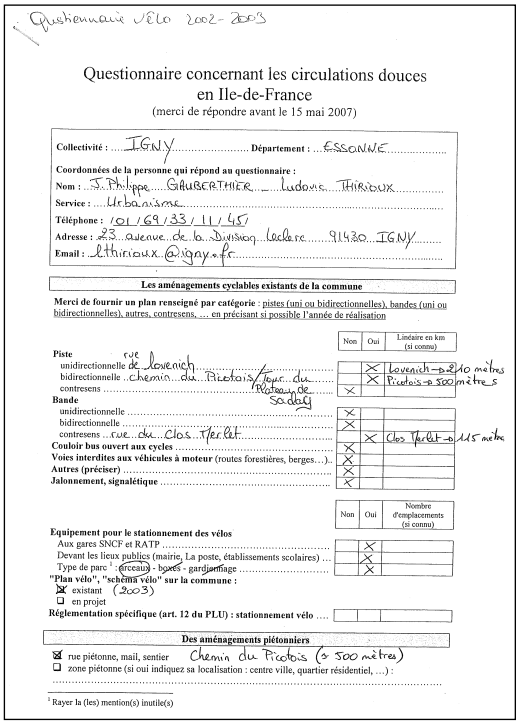
Let's switch now to a completely different setting. I worked for almost two years with a small company that provides itineraries for cyclists online. One of the reasons the founder of this startup came to me concerned the way they gathered information regarding bike facilities. They used to produce these geographical data using *OpenStreetMap*, mostly by their own and with the help of a few volunteers. But they wanted to ask local administrations for their data, in order to import them directly into the database that fed their algorithm.

I won't explain in details this study, but what mainly interested me was the position of the manager of the company, who, at first, really took the fact that local authorities had data about bike facilities for granted. This is actually another side of data positivism. A positivism one can find in most calls for open government data, and that assumes that data are already-there entities, just waiting for their "release."

heterogenous information, unsuitable data, and false numbers

In order to help the startup, I visited a few French city halls in search of data. And the founders of the company and I progressively realized that most of information regarding bike facilities could not be considered as data. It existed in heterogenous forms, and some of it wouldn't even let any written traces. Moreover, when such things as data were spotted, they were far from being ready-to-use for the algorithm. Some of them weren't updated regularly enough, some others displayed what Lampland (2010) calls « false numbers » (for instance the dates of the beginning and the end of street works).





Questionnaire Vélo 2002-2003

Questionnaire concernant les circulations douces
en Ile-de-France
(merci de répondre avant le 15 mai 2007)

Collectivité : IGNY Département : ESSONNE

Coordonnées de la personne qui répond au questionnaire :
Nom : S. Philippe GAUBERTHIER Ludovic THIRIOUX
Service : Urbanisme
Téléphone : 01/69/33/11/45
Adresse : 23 avenue de la Division Leclerc 91430 IGNY
Email : lthirioux@igny.fr

Les aménagements cyclables existants de la commune

Merci de fournir un plan renseigné par catégorie : pistes (uni ou bidirectionnelles), bandes (uni ou bidirectionnelles), autres, contresens, ... en précisant si possible l'année de réalisation

	Non	Oui	Linéaire en km (si connu)
Piste			
unidirectionnelle rue de Lovenich		X	Lovenich → 210 mètres
bidirectionnelle Chemin du Picotois/Sente du Picotois		X	Picotois → 500 mètres
contresens	X		
Bande			
unidirectionnelle Rotonde de Saclay	X		
bidirectionnelle	X		
contresens rue du Clos Perlet		X	Clos Perlet → 115 mètres
Couloir bus ouvert aux cycles	X		
Voies interdites aux véhicules à moteur (routes forestières, berges...)	X		
Autres (préciser)	X		
Jalonnement, signalétique	X		

	Non	Oui	Nombre d'emplacements (si connu)
Equipement pour le stationnement des vélos			
Aux gares SNCF et RATP		X	
Devant les lieux publics (mairie, La poste, établissements scolaires) ...		X	
Type de parc [1] : arceaux - boxes - gardiennage		X	

"Plan vélo", "schéma vélo" sur la commune :
☒ existant (2003)
☐ en projet

Réglementation spécifique (art. 12 du PLU) : stationnement vélo ....

Des aménagements piétonniers

☒ rue piétonne, mail, sentier Chemin du Picotois (≈ 500 mètres)
☐ zone piétonne (si oui indiquez sa localisation : centre ville, quartier résidentiel, ...) :

[1] Rayer la (les) mention(s) inutile(s)

This exploration, followed by my explanations, allowed two things. First it helped the manager of the company understand that such messy informational practices were not scandalous. He progressively accepted that there were "good organizational reasons" (Garfinkel & Bittner, 1967) for what he saw at first as bad data or not data at all. And second, it helped him to acknowledge that the very production of reliable "good" data, especially with *OpenStreetMap*, could be an asset his company and its product.

ecology of visible and invisible work

Of course, these two very different studies were mainly aimed at surfacing invisible data work and to bring data workers to light. Yet, they also prompt to investigate not only the invisibility of work in itself, but what Star and Strauss called the ecology of visible and invisible (Star & Strauss, 1999). It notably urges to study the process and the reasons of the invisibilisation of data labor, its perimeter (to whom is it made invisible), and it also invites to explore the variety of configurations, and to investigate situations where conversely, data labor is made visible (and again, to whom and in which conditions).

when is data? what counts as work?

Yet, we should understand that the ecology of data labor visibility and invisibility is not only the outcome of specific organizational configurations, such as Amazon Mechanical Turk for instance, which performs, as Lilly Irani (2013) showed, a clear distinction between creative visible work and menial invisible work. What counts or does not count as work is also entangled in the very definitions of data itself.

As Sabina notably put it (Leonelli, 2015), since the same objects may or may not be functioning as data regarding the situations, data is better understood as a "relational category." Thus, instead of asking *what is data*, we'd better ask, *when and how is data?* This means that, to borrow Dorothy Smith's vocabulary, the ontologies of data are situated and enacted (Smith, 1974). What I would like to point today is that these enacted ontologies of data always carry a more or less explicit moral distribution of work.

In the bank, good and reliable data are supposed to circulate without workers, who are only present to spot bad data. For the company that provides cycling itineraries, what counts as data is what can be directly imported to the cycling algorithm. Everything that explicitly would have taken translation, or adjustments, was at first not considered as data at all.

We could find a lot of other sites of research to explore these issues. For instance, we can witness around open data policies a lot of *data shaming*. Blog posts or journal articles stigmatize datasets for their bad quality, or for not being data at all.

> 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, Bad Data, http://okfnlabs.org/bad-data*

What do such claims express, actually? When saying, We don't want non-consistent categories, we want only .csv files, and no Excel spreadsheets full of merged cells and colored rows, above all we do not want pdf (pdf is evil). they also perform a moral economy of data labor. They say, "It's your job not ours. We don't want to reprocess your data, we don't want to extract the letters and the numbers from your pdf files." Basically, they say, "Give us raw data that are both unmodified and already machine readable. We don't want to know what it takes."

Yet, we just saw that this is not the only way to articulate what counts as data and what counts as work. On the contrary, we can find places such as some *OpenStreetMap* communities where everybody knows and claims that data always comes with work. Even if still exploratory, I think such contrasted positions show that the identification of what counts as work is a heuristic preoccupation that can help a better understanding of what counts as data.

references

Agar, J. (2003). *The government machine. A revolutionary history of the computer.* Cambridge: MIT Press.

Beniger, J. R. (1986). *The control revolution. Technological and economic origins of the information society*. Cambridge: Harvard University Press.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge: MIT Press.

Denis, J., & Pontille, D. (2012). Workers of Writing, Materials of Information. *Revue d'anthropologie des connaissances*, 6(1).

Gardey, D. (2008). Écrire, calculer, classer. Comment une révolution de papier a transformé les sociétés contemporaines (1800-1940). Paris: La Découverte.

Garfinkel, H., & Bittner, E. (1967). 'Good' organizational reasons for 'bad' clinic records. In H. Garfinkel, *Studies in ethnnomethodology* (pp. 186–207).

Goody, J. (1986). *The logic of writing and the organization of society*. Cambridge: Cambridge University Press.

Irani, L. (2013). The cultural work of microwork. *New Media & Society*, 17(5), 720–739.

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

Latour, B. (1987). *Science in action. How to follow scientists and engineers through society*. Harvard, Harvard University Press.

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

Smith, D. E. (1974). The social construction of documentary reality. *Sociological Inquiry*, 44(4), 257–268.

Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, 43(3), 377–391.

Star, S.L. & 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.

Suchman, L. (1995). Making work visible. *Communications of the ACM*, 38(9), 56–64.

Yates, J. (1989). *Control through communication: The rise of system in american management*. Baltimore: Johns Hopkins University Press.