Data cleaners for pristine datasets: visibility and invisibility of data processors in social science

Introduction

Statistics' increasing significance in modern government since the 19th century (Porter 1986; Desrosières 2010), combined with the development of data collection and processing techniques for large-scale datasets (Campbell-Kelly 1990) have contributed to the importance of survey research in social science since the second half of the 20th century. Notable examples of such studies are the Roper Poll, the General Social Survey, the American National Election Studies, and the Current Population Survey (Converse 2009). Concurrently, dedicated institutions were created to archive and disseminate such datasets for secondary analysis. Examples of these social science data archives (Shankar, Eschenfelder, and Downey 2016) include the Inter-University Consortium for Political and Social Research (), created in 1962, the UK Data Archive in 1967, the Norwegian Data Archive in 1971, and the Consortium of European Social Science Data Archives (CESSDA) in 1976. Similar to cyberinfrastructures (Atkins 2003) or knowledge infrastructures (Edwards 2010), these institutions organize the large-scale dissemination and long-term preservation of data (Edwards et al. 2009; Ribes and Finholt 2009) in social science.

To fulfil these goals, these institutions must cope with all the imperfections inherent to research datasets, either resulting from researchers' mistakes or idiosyncrasies, or from coordination problems between the different actors involved in data collection and analysis (such as market research firms, research assistants, etc.). To compensate for such inherent flaws, the data archive at the center of this article hires dedicated workers called "data processors," or "data curators," whose work consists of reformatting and "cleaning" the deposited datasets before archiving them for use by researchers for secondary analysis.

In this article, I investigate how these data processors work, how they contribute to data sharing, and how their work is simultaneously visible and invisible to other parties. I base this analysis on ethnographic fieldwork conducted at a major US data archive that specializes in quantitative social science data[1]. This empirical work consisted of participant observation at the processing unit of this data archive, where I worked as a part-time intern for six months in 2014. Through this position, I received the same training as newly-hired data processors: I worked under the supervision of a senior processor, had my own work station, and learned by processing existing datasets. This participant observation was complemented by fifteen semi-structured interviews, conducted in 2014, with the eight data processors working at the processing unit at the time, and with seven employees having different roles across the archive: the director of the archive, the director of acquisition, the process improvement specialist, a metadata librarian, and various managers of the processing unit.

Based on this investigation, this article presents two main results. First, it contributes to the literature on invisible technicians in scientific work (Barley and Bechky 1994; Timmermans 2003) by showing that the same organization of labor that makes technical work *invisible* to other parties involved in the research process can also make it completely *visible* to others, such as managers and other employees of the archive. The data processors studied here, in addition to having all their (few) interactions outside the archive strictly framed, also have to make all their processing activity on datasets completely explicit to their peers and managers inside the archive, who will make sure they process datasets according to procedures and eventually produce a

[1] The institution and all the staff members have been anonymized in this article.

standardized dataset. Second, this article contributes to the social study of scientific data sharing. If the conception that researchers have of what counts as data has a direct influence on their data sharing practices (Wynholds et al. 2012; Wallis, Rolando, and Borgman 2013; Borgman 2015), it is similarly the case for data archive. I show that the organization of data processing labor at the data archive directly stems from archive managers' conception of a valid dataset—that is, a dataset that must look "pristine" at the end of its processing. This goal of achieving "pristineness" for datasets is translated in an organization of processing work that makes the processors' activity invisible outside the archive, while making it completely visible inside the archive to peers and managers. This conception of data, and how it shapes data processing, eventually obscures the processors' social function of intermediary and their contribution to data sharing.

This article first reviews existing research on invisible work in science and on scientific data sharing. Then it details the data processing pipeline from the deposit of a dataset to its publication on the archive website. Next, the article explains how the work procedures of this pipeline keep the work of processors invisible to those outside the archive: it describes the warnings in the processing manual against being too creative; all the procedures that processors must follow in the rare cases when they have to contact depositors; and the failed attempt to make internal data processing notes public. In the third part, the article shows that despite different ways of working, processors must follow guidelines that make their work completely explicit to their colleagues, in preparation for the final quality control before publication. This visibility was implemented at a time when the institution wanted to fight idiosyncratic ways of working, but now it lowers the incentive to innovate at the workplace. In the fourth part, I

critically examine the notion of pristineness and show how it perpetuates a misleading conception of data as "raw." This article concludes by showing that as long as the institution will rely on such erroneous conception of data to organize data processing, it will reproduce the invisibility of data processing staff instead of acknowledging how critical their intermediary function is to data sharing.

Technical work in science and its role towards data sharing

Social studies of science have shown that despite its objective pretensions, scientific work depends in large part on social contingency (Lynch 1982) and tacit knowledge (Collins 1974), parameters close to "magic" (Cambrosio and Keating 1988). These instances of "shop work and shop talk" (Garfinkel, Lynch, and Livingstone 1981) are not simply what happens between formal procedures, they constitute the "craft" necessary to manipulate scientific instruments (deSolla Price 1984). Following this impulse, many studies have precisely shown the important contribution of technicians to the research process: beyond preparing tools and equipment, their work requires interpretation and creativity (Shapin 1989), up to the point of developing skills similar to scientists (Timmermans 2003). However, this skillset is almost never enough to revert the hierarchy between the different layers of professional status in science (Barley and Bechky 1994), and their work remains invisible in the history of science.

Data processors working in data archives are among the invisible workers of science: their activity, between the deposit and reuse of datasets, remains largely unknown from other actors involved in the social science research process. However, their work differs from laboratory technicians inasmuch as they do not work at the "empirical interface" (Barley and Bechky 1994) of scientific work. Rather, they are situated in the "intermediary zone" of the archive, between

data collection and future secondary analysis. They are not part of any data collection, nor data reuse, and only rarely are they in direct contact with researchers. Their work is therefore closer to data managers (Baker and Millerand 2010; Millerand 2012; Dagiral and Peerbaye 2012), who perform tasks such as post-hoc data entry, verification, or metadata writing.

This article contributes to this scholarship by pushing further the relation between the concept of invisibility and visibility of technical work in science. The relational perspective promoted by Star and Ruhleder (1996) to study information infrastructures (Bowker et al. 2010) has shown that what counts as "work" varies depending on the indicator: slavery is the most extreme example of work made invisible (Star and Strauss 1999), while the entrance of domestic work in national statistics is an example of a change of classification that makes work visible (Bowker and Star 1999). The case study of processing in data archives shows that technical scientific work can *simultaneously* be invisible and visible. Data processing is organized, on the first hand, around procedures that make processors' work completely invisible *outside* the archive: strict work procedures are applied in the data archive to frame any contacts from processors with researchers outside the archive, as well as removing any traces of processors' agency from the final published dataset. On the other hand, the same procedures make the work of processors completely visible to other processors and the management team, who can then verify that processors' work in a similar fashion and produce similar outputs.

The second contribution of this article is to explain how the data archive justifies the necessity of this processing work—something that I relate to the conception of "pristineness" promoted at the archive and used to judge if datasets are ready to be shared. Existing research on data sharing has

extensively studied the relation between the multiple (and often competing) definitions that researchers have of what counts as data and their data sharing practices (Wynholds et al. 2012; Wallis, Rolando, and Borgman 2013). By asking "when are data?" Borgman (2015) shows that definitions of data are inherently related to institutional contexts, dynamics of epistemic communities, and different uses of data in the research process—all resulting in highly contrasted data sharing practices between sciences, social sciences, and humanities. I extend here this research on what "counts as data" and apply it to the context of data archive.

The data processing unit studied here relies on the criteria of "pristineness" to assess if a dataset, after its processing, possesses the required characteristics to be published for archiving or reuse. The work of data processing therefore revolves around this organizing principle: keeping the work of data processors inside the archive, and conducting peer review of their processing work, both intended to remove all traces of data processing to restore a presumed original state of the data. However, this notion of "pristine data" can very easily be deconstructed: echoing recent debates on the oxymoronic nature of the concept of "raw data" (Bowker 2005) (Gitelman 2013) (Denis and Goëta 2017), presenting data as a pristine product conceals all the work needed to process and prepare such data for publication by the archive. Using such term, and translating it into the work process described below, therefore reduces the importance of the role of data processors in making data circulate between production and reuse (Leonelli 2016), and eventually keep processors' contribution to data sharing invisible.

1. The data processing pipeline: making data "pristine"

The table below formalizes the current journey of a dataset within the data archive, from the moment it is deposited by a principal investigator (PI) to its final publication on the website. Each step engages different actors, techniques, and actions, summarized below:

Table 1. Reconstitution of the "pipeline" for data processing by the author

The deposit step (Step 1) is the entrance point of the data inside the archive, typically occurring through researchers depositing datasets they want to archive. After an archive manager briefly reviews the data to see if it fits the selection criteria of the institution, the study is dispatched (Step 2) to a processor. They receive their assignment by a notification on the dedicated internal secure workspace through internal email. The data processing can start.

The Review and Process step (Step 3) aims to identify and to "fix" the problems that datasets contain. Called "undocumented" or "wild code," these problems consist of irregularities and formatting issues, including "any value that is not a valid code or is not properly documented."[2] For example, in a dataset (say, in SPSS), some variable labels or value labels might be missing, incomplete, or containing suspicious characters. Many discrepancies can quickly be "eyeballed," such as a missing document that should have been submitted along with the datasets, or the discrepancies between the numbers of questions in the questionnaires versus in the datasets. This "cleaning" step is also partially automated through custom-made scripts, which provide a faster

[2] Extract from the processing manual.

assessment of the state of the dataset (providing, for example, a summary of "unlabeled variables" or "variables without any valid cases") and flag areas of the dataset that require fixing.

Once this step is done, the next part of data processing (Step 5, Metadata and Formatting)[3] is to prepare the data for online publication on the website, through the creation of metadata and documentation. For the former: once a processing staff member has "cleaned" the data of its potential flaws, they have to generate versions of the dataset in the multiple statistical packages (SPSS, SAS, STATA, etc.) that the institution provides. For the latter: the archive maintains a thesaurus of social science topics and domains used to categorize datasets, which helps their retrieval on the website. Processors therefore have to enter these metadata for the dataset they are processing, such as the scope of the study, the methodology, related publications, etc. Processors also create codebooks and edit documents (such as questionnaires) to be published online along with the datasets (cf. Step 7 on the Table 1).

Despite their apparent objectivity (Busch 2013, p. 68), standards are "agreed-upon rules" that emerge from specific communities of practice (Bowker and Star 1999, p. 13): they translate specific values into action, and are "morally charged matter" (Busch 2013, p. 22). The institution puts at the center of its action the necessity to "clean" data in order to allow its further circulation and reuse in social science. Data processing, organized through the formalized pipeline just described, is the means to achieve "pristineness" of data, deemed essential for data reuse. As a

[3] I develop later in the article the optional Step 4 of contacting the original depositor.

manager at the processing unit puts it: “We want everything to be right, and everything to read properly [...] Trying to get that, so that the future users when they get [the data], they get everything in a pristine manner.” “Pristineness” is the core principle that shapes the multiple procedures to process data at the archive. While I present a critical reading of this notion in the fourth section of this article, I describe in the next two parts how this goal of “pristineness” is achieved by rendering invisible data processors’ work outside the archive while making completely visible their work inside the archive.

2. Making processing invisible outside the archive

This section describes the three steps that render data processors and their work invisible to outside users: (1) With rules preventing processors from leaving comments about the state of data on which they worked; (2) With rules limiting processors’ contact with users outside the archive; (3) Through lessons learned from the aborted attempt to disclose the “processing history file” (explained below) to future re-users of data.

2.1 “Don’t get carried away, though”

Through their deep engagement with the structure of the data, as well as their access to all the materials used to conduct the study, processors gain an insider’s view on the methodology of a study—and on its potential flaws. They can develop, with time, knowledge of research methodology and can assess the quality of the study—e.g. the quality of the questions or the types of variables chosen—beyond just spotting and fixing basic design flaws. As one processor mentions: “Sometimes I look at the stuff that comes to us and I'm like, ‘Jesus, this is terrible research!’” Another processor mentions: “Every once in a while I'm like, ‘That's not how you ask a question.’” However, processors are fully aware that methodological critique is not their

job and that they cannot voice such concerns in the final product. They adopt in their work a forced blindness, in which they only look at the structure of the data, and not at the content, the design, or the results of the study. As Caroline says[4]:

> What are we supposed to do at that point? Am I really going to put a processing note saying, "User should note that this questionnaire is flawed." No. We wouldn't do that.

The way the processing pipeline is designed does not leave any space for writing a note on the quality of the study, which might be contested by the management during the final quality control.

In case they forget their position, the processing manual clearly reminds them: "Don't get carried away, though. Our job is to make the data available to the research community, not to critique the original researcher's methods and conclusions."[5] This is a direct reminder that data processing staff are not researchers, and that their action has to stay at the level of the "cleaning" of datasets. The final check, which we describe later, applies here to make sure the processors direct their actions at the level of "cosmetic" changes, and do not venture into criticizing the design of the study or the results.

[4] All the names in this article are pseudonyms.

[5] Extract from the processing manual.

2.2. Breaching the invisibility: contacting the Principal Investigator

As usual with infrastructures, invisible work becomes visible when something goes wrong or breaks (Bowker and Star 1999). This is the case when relying only on the documentation provided by the original Principal Investigator (PI) cannot solve a problem in the data. Problems of such magnitude are among the few instances when processors are allowed to breach the boundary of the data archive and to enter into contact with researchers.

A study I was processing had such a major problem: the number of variables in the SPSS files containing the data was very different from the number of questions written in the questionnaire. Did such discrepancy come from the fact that not all the questions present in the questionnaire were asked? Or were they all asked, but columns in the datasets disappeared due to conversion problems, or other issues? As this discrepancy was not explained anywhere in the accompanying documents deposited with the dataset, my mentor and I came to the conclusion that I had to contact the PI to ask him/her directly (cf. Step 4 on Table 1).

Contacting the PIs at this occasion represents processors' sole contact with the outside world of research. This excursion outside the archive is considered only as a last resort, and like other steps along the pipeline, it must go through a strict procedure, clearly explained in the processing manual. The processor contacts the PI by email, pastes a standard email, adds the title and deposit number of the study, and attaches a spreadsheet that "lists the variable name, the dataset name (if the study has multiple parts) and the problem issue." With these information and

documents, the PI is supposed to "sort by dataset, variable name, type of problem, etc." and to "put the correct information in the excel file and return it."[6]

Such intervention may be unsuccessful, because PIs regularly do not respond to such requests[7]. In this case, the processor will add a mention on the first page of the final public version of the data, explaining the problems that remain in the dataset. The first breach of invisibility of the processor through contacting the PI, when unsuccessful, results in a second breach in the documentation published on the website along with the data. A disclaimer warns future users problems were identified by the data processor that could not be fixed due to the absence of response from the PI. This message re-assigns the responsibility for the poor quality of a study to the PI and exonerates both the processors and the institution.

2.3. The risk of disclosing too much

While invisibility can be breached when something goes wrong in processing, an aborted attempt a couple of years ago to disclose publicly and systemically the details of internal data processing similarly revealed the institutional logic behind keeping data processing invisible. When they work, processors continually update a text file called "processing history file" (simply called "ph file"), where they keep track of all the actions they have performed on data. It aggregates the logs of the different scripts performed on the data, as well as notes written by the processors to

[6] Extract from the processing manual

[7] I left the archive before receiving an answer from the PI concerning this problem and therefore never learned the cause.

explain the remedial actions taken. It constitutes an internal text file that acts as a notebook. However, its state changes as the processors go further down the pipeline and eventually constitutes the basis for a second external document, simply called the "processing note." Whereas the ph file remains an internal document, the processing note is inserted at the beginning of each published version of the dataset.

This transition from internal document to public note is a summary of all actions the processor took. The initial processing history file is a document of several pages, detailing the number of variables and cases, containing the logs of the scripts run on the data, potential emails with PIs, etc. It aims to document exhaustively the changes processors performed on the data, and is mostly intended for internal reviewers who can then track and understand what happened. At the other end, the public "processing note" is a much shorter document, usually a couple of sentences long, and contains only the information necessary to use the data, for example if the data are weighted or not.

A couple of years ago, the option to publish the whole processing history file, and not only its shorter public version as processing note, was discussed at the management level of the archive. The rationale was mostly to increase the transparency surrounding how the institution works on the data deposited. However, this option was not taken, eventually, for fear that it would breach the guarantee of standardization of its data. As mentioned earlier, the archive's emphasis on standardization concerns the final product, rather than how it is achieved. Showing all the variations in processing and making public all the ways processors "work differently" would, it was feared, undermine the impression of standardized internal processing. A manager of the

processing unit remembers worrying that such disclosure would create more questions from both depositors and re-users about the internal data processing: “How [do] you do [such disclosure] without having people ask questions, especially when you can write the code and do the processing in different ways?” By keeping the slight variability between processing styles inside the archive, instead of making them visible to the future re-users, the archive reinforces the impression—for both depositers and re-users—that data at the institution all follows the same processing guidelines.

We saw in this part how the data processing pipeline erases all traces of processing in the final product, which simultaneously makes the data processors invisible to researchers outside. In the next section, we see how the goal of producing “pristine” data results paradoxically in the visibility of processing work inside the data archive: the high standardization and the multiple verification procedures exist to make the work of processors completely explicit to each other for quality control.

3. Making processing visible inside the archive

The same standards and verification procedures described above are in place to make explicit the idiosyncratic ways processors work, thereby reducing potential variations in the final output. I first describe how processors develop different ways of working, mostly to fight routine and to add challenge in a very standardized environment. However, I then show that such idiosyncratic ways of processing are tolerated only on the condition that they are made completely explicit and visible to the other workers: if processing practices do not follow such constraints, they are actively discarded and labelled as “folklore.” Third, I show how such constraint of visibility is

implemented through record keeping and final quality control of their work, which results in the curtailing of the agency of processors on the pipeline.

3.1. "Everybody processes differently"

Despite the firm commitment of the data archive management team to implement formalized routine, data processing includes room for agency and initiative, and processors develop a wide range of expertise and knowledge beyond the strict division of labor that defines their work.

A first striking fact that contrasts with the idea of a homogeneous work practice is the consensus within the archive that everybody processes differently. Steps of data processing are presented as tightly entangled and to be conducted following a strict order (e.g. running scripts before analyzing the structure of data), yet data processors routinely alter the order. The step of writing metadata and documentation can, for example, be started almost at any time in the pipeline, as it mostly concerns the topic of the original deposited study, and can easily be updated later to reflect the result of the processing work. Those not having a particular interest in metadata can postpone it to a later stage, when it can no longer be avoided.

Data processing entails a very distant reading of datasets, and any incursion inside the data takes the form of research and discovery. Looking at the results of a study is not strictly necessary for processing the dataset, but it offers another way of fighting the boredom while manipulating the materials that are at the center of the work. As Taylor explains:

> It's my opinion that you can't work eight hours straight [...] So with those studies, though, you can just read the study and to me that's really interesting data. As I'm reading through the study, I'm like, "Oh I wonder if I crossed half like those two, what do these have in common?" That gets fun and then that's a way where you can take a break from the processing.

Star (2002) famously invited researchers of infrastructures to study boring things to reveal how exciting they are. Data processors similarly develop a multiplicity of strategies to make their work more exciting and to fight the inherent boredom of such repetitive and highly standardized work. The management team is fully aware of all the adaptations in data processing and tolerates them. As a manager says: “Some people like to do things one way, others like to do it another; [...] we don't dictate how that happens, but people do their preferences.” However, it is only under this condition of being made completely explicit that idiosyncratic ways of data processing can be tolerated. If such obligation of visibility of the work process is not met, the institution will discard such knowledge as merely “lore” and “tales,” as opposed to explicit and reproducible work procedures.

3.2. Processing guidelines vs. “folklore”

In the early 2000s, the archive hired an external consultant to study the different ways processors work on data, in order to streamline the process. The key issues presented to the specialist as urgent were the lack of a single reference way of processing, coupled with highly singular ways of working. Through this process, he could appreciate the high level of idiosyncrasies in how each processor worked at the time:

> because the only thing we really knew about this task over here was it was something magic that Janet did. [...] Over here, there was something called, “throw the boiling frog in the boiling cauldron” or something like that, some bit of sorcery and all we knew is that it was something that Sheila did, and efforts of asking Janet and Sheila for detail in the past have not been successful, but it was clear that they had to be done because things would fail otherwise.

Retrospectively, the specialist describes his action at the institution as an effort to switch from what he calls “lore” and “tales” of processors—singular ways of processing data resulting from years of practice—to standardized procedures made explicit and shared by all.

His method consisted of interviewing and observing processors at work, asking them to describe their activity, and comparing these accounts between different employees. What came out of this series of interviews and observations was a diagram of the different steps that constitute data processing, as the basis of the pipeline currently followed at the archive (reproduced in the processing manual available to processors, and reconstituted in table 1). These lore or tales, acquired through years of experiences and that remained unknown to other workers, were replaced with clearly identified, publicly known, and chronologically ranked procedures that eventually became explicit checklists to follow. As a consequence, the verification procedures currently in place in the archive exist to make data processors constantly explain the decisions they took in their work, to make sure their agency does not feed new “lore” and “tales” about competing ways of processing. These procedures combine a specific document, the processing

history file, with a peer review of all the outputs from data processing called quality check (cf. Step 6 on the Table 1).

3.3. Complete visibility as disincentive to innovate

As we saw earlier, data processors record all their actions on data in the processing history file. Once they have reached the end of the Stage 5 ("metadata and formatting"), they send all the documents constituting the study (processed datasets, reformatted documentation and metadata) as well as the processing history file for inspection. This penultimate step, called "quality check," is essentially a peer review of the work done. Both a fellow data processor and a member of the management team alternately go through all the outputs of processing work, and ask specific questions: Are some "wild codes" still unfixed? Are the codebooks and metadata following the archive's templates and standards for publication? Are all the documents present in the folder ready for publication? Depending on the answers, a back and forth between reviewers and processors can occur to ask for further questions about one stage, and even to send back the work for further processing.

This final verification reveals the standardization of work practices resulting from making all the processors' ways of working completely visible. The final quality check illustrates the obligation of visibility that comes with this work: all the stages of data processing need to be made completely explicit and visible to the reviewers. If something is missing or is not clear, the reviewer will ask the processors to elucidate how one stage of the pipeline was conducted. This complete visibility results in the elimination of any traces of agency from the processors. We mentioned already that each processor works slightly differently, and that it is tolerated as long

as their work remains close to the standardized procedures. However, this high level of verification also acts as a strong incentive for processors to rely on tools and processes that are common practice in the archive. Consequently, it acts as disincentive to develop innovative ways of processing data, as they would not fit easily in the highly standardized existing work procedures.

Taylor reveals in an interview how he once found a more efficient way to simultaneously delete a large series of characters by using a technique not in the processing manual and a programming language other than those commonly used. However, he quickly realized how this innovation would not fit anywhere in the processing pipeline: “there's no way to record what I did, so I had to just write a paragraph explaining what I did.” This innovative way of solving a problem is tolerated, as long as it is made explicit in the processing history file; however, the verification procedures in place do not leave the space for this initiative to become standard procedure, and will eventually not go beyond the status of temporary “trick.”

I have shown how procedures make the work of data processors completely invisible to researchers outside the archive, yet completely visible to their colleagues and managers inside the archive. In the next section, I show that this double constraint stems directly from the conception of data promoted by the institution: the archive takes as mission to deliver “pristine data,” which incidentally renders invisible the contributions of processors to data sharing.

4. Cleaning Data Twice: Social Use and Critique of Data "Pristineness"

At the end of the pipeline described above, datasets are judged ready for publication when they possess two qualities, assessed during the final quality control. First, managers and peers verify that datasets are cleaned of all the flaws, irregularities, and other idiosyncrasies from the original data producers. This is the first level of cleaning, completed in the part 3 of the pipeline ("Review and Process," cf. figure 1): processors here remove elements of the original context and ensure that datasets can be used in future contexts without biases or difficulties. The second and equally important cleaning concerns the work of processors themselves. Datasets are deemed ready when no traces of their cleaning and repair remain in the final dataset—something completed during the stage 5 of the pipeline ("Metadata and Formatting"), where data processors make sure that the final output follows standardized formatting. Similarly, by verifying that processors don't get "carried away" in criticizing a study, and that all outputs are similar despite singular ways of working, other stages in the pipeline aim to erase any modifications of the original dataset. Processors therefore apply the same guidelines and procedures, first, to the mistakes and flaws from previous studies, then to their own work, so that they also remove all traces of their own cleaning. When they are sent for quality control, datasets have theoretically been cleaned *twice*—of loose ends from the original production, but also of any residue from their own processing. It is only under these conditions that datasets can eventually be considered pristine, and therefore ready for publication.

Latour and Woolgar (1979) have shown that a similar process happens during the creation of scientific facts. Scientists remove competing and uncertain statements (the "modalities," cf.

p. 81) to eventually position their interpretation as an incontestable fact. A parallel process happens in the data archive, but for a different purpose. The goal here is not to construct scientific facts, but to deliver standardized datasets that are pristine. This is first achieved by removing the flaws from previous studies; second, by restructuring the dataset to fit the archive's template. As the processing pipeline is designed to make sure that no traces of this double cleaning appear in the final datasets, the data eventually appear at the end of this process as "raw" again (Denis and Goëta 2017).

Pristineness occupies a key social function in the data archive: it acts as the leading principle that is implemented in the various stages of data processing (the "pipeline"), and in the work procedures and quality checks. However, emphasizing data as "pristine" runs the risk of perpetuating a misleading conception of data that overlooks all the various stages of production and formatting that occurred before a dataset is judged valid. Many commentators have shown how the recent trend of big data research can reproduce a positivist conception of data as central unique arbiter of truth for valid scientific result (Anderson 2008; Hey, Tolle, and Tansley 2009). Following up on critical research that has largely debunked such claims (Bowker 2005; boyd and Crawford 2012; Gitelman 2013), data archives are in a unique position to show that data never comes as raw, pristine, or ready to use, but that multiple interventions are always needed before data can be reused.

Moreover, emphasizing a presupposed pristineness of datasets obscures all the intermediary work that is necessary to make data sharable and reusable. Scholarship on data sharing has shown that data do not circulate by themselves: they must be prepared, even "packaged"

(Leonelli 2016), to circulate among different parties and to "jump" between different contexts (Downey 2014). On the opposite, when the notion of pristineness is used to characterize the final output of the data archive, it operates a reduction of the very specificity and richness of data archives—that is, to hire dedicated employees who enhance the data they receive before reuse. As long as the institution commits to such a conception of data, it will continue to conceal all the information labor (Downey 2014) provided by its processors, and will thereby keep them invisible, instead of recognizing their essential contribution to the circulation of data in social science.

Conclusion

A rich scholarship in the social study of science has expanded the scope of actors that play a key role in the construction of scientific facts. Starting in the 1980s, multiple studies have specifically acknowledged the role that technicians play at the empirical interface of science and in the larger division of scientific labor (Barley and Bechky 1994; Timmermans 2003). This historical research provided the basis for the investigation of technicians in information infrastructures (Star and Ruhleder 1996; Bowker et al. 2010) focusing on the data managers who pre-process and prepare data for analysis (Baker and Millerand 2010; Millerand 2012; Dagiral and Peerbaye 2012).

The ethnographic study presented in this article extends this study of invisible workers in science by focusing on one type of technicians: the employees in the processing unit of a social science data archive who work at the intermediary level between the deposit of datasets and their publication for archiving and reuse. By focusing on these data processors, who do not directly work at the empirical interface of scientific work but who are still involved in the circulation and

reuse of data, this article makes two contributions. First, it shows that technicians may be invisible to some but are made completely visible to others, such as colleagues and managers in their workplace. Invisibility and visibility are implemented through the same work procedures, and both stem from the goal of the institution to publish "pristine" datasets. Second, this study contributes to the social study of data sharing by detailing the work of data cleaning and reformatting that happens at the intermediary zone between data production and reuse. Doing so expands studies that have detailed the conditions and practices of data sharing (Wynholds et al. 2012; Wallis, Rolando, and Borgman 2013; Borgman 2015), but it emphasizes the preparation required to circulate data among different contexts (Downey 2014; Leonelli 2016).

Finally, investigating these invisible processors shows the value of a critical study of data in science. It is paradoxical that the data archive studied here promotes a conception of data as "pristine" while its managers know that there is no such thing: the data processors employed there do much more than reverting to a presumed original state of "raw" data. They actively improve and enhance the data they work on, and the datasets that come out of the data archive are of better quality than when they were submitted. By "going backstage" (Star and Strauss 1999) and observing and practicing all the "boring" aspects (Star 2002) involved in processing data, this research takes part in the critical deconstruction, triggered recently by the rise of "big data" research, of data as an immediate commodity (Bowker 2005; Gitelman 2013). By putting this intermediary work and all the multiple stages of "cooking" data to the fore, data archives could be part of this debate, as they are the living proof that data processing is crucial to the circulation of data in science.

The motivation for data archives to make the work of their data processors more visible could come from competition emerging in the data archiving scene. The rise of "big data" science and of disciplines such as digital humanities challenge the traditional properties and missions of knowledge infrastructures (Edwards et al. 2013; Karasti et al. 2016). The past few years have seen the rise of web-based digital platforms that occupy key functions traditionally fulfilled by infrastructures, directly competing with data archives (Plantin, Lagoze, and Edwards 2018). Exemplified by the service Figshare.com, such entities promote self-deposit and quick release of deposited data over a thorough but slower processing of research data detailed in this article. The rise of these entities, which do not provide the same guarantee in terms of curation, long-term archiving, and metadata, could force data archives to openly embrace the value added by their data processors, hence making their contribution more visible to the world of research.

Acknowledgement

I would like to thank the editors of the journal and the two anonymous reviewers; Paul N. Edwards, Carl Lagoze, and Christian Sandvig for their guidance during the research that led to this article; Meghanne Barker, Nick Couldry, Justin Joque, and Florence Millerand for their generous comments on earlier versions of this article.

The research was presented at the Université de Lille, Conservatoire des Arts et Métiers, SciencesPo Medialab, and at the Stevens Institute of Technology (for the conference Maintainers II: Labor, Technology, and Social Orders, April 6-9, 2017). I would like to thank respectively Marta Severo & Eric Dagiral, Manuel Zacklad & Lucile Desmoulins, Paul Girard & Ian Gray,

and Andrew Russell & Lee Vinsel, as well as audience members whose comments that improved the quality of this article.

Funding

This article results from research conducted during the project: "New directions in the Study of Infrastructures," funded by the MCubed program, University of Michigan , 2013-2014 (Principal Investigators: Prof Paul Edwards, Dr Carl Lagoze, Prof Christian Sandvig).

Author biography

Jean-Christophe Plantin is Assistant Professor at the London School of Economics and Political Science, Department of Media & Communications. He investigates the civic use of mapping platforms, the collaborative challenges in big data science, and the evolution of knowledge infrastructures. His research has been funded by the Alfred P. Sloan Foundation, the Gordon and Betty Moore Foundation, the European Regional Development Fund, and the University of Michigan MCubed Program. Jean-Christophe is currently writing a monograph on the infrastructural evolution of digital platforms.

Reference list

Anderson, Chris. 2008. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” WIRED. June 23, 2008. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Atkins, Daniel. 2003. “Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure,” January. http://arizona.openrepository.com/arizona/handle/10150/106224.

Baker, Karen, and Florence Millerand. 2010. “Infrastructure Ecology: Challenges in Achieving Data Sharing.” In *Collaboration in the New Life Sciences*, edited by John N. Parker, 111–38. Ashgate Publishing Limited.

Barley, Stephen, and Beth Bechky. 1994. “In the Backrooms of Science: The Work of Technicians in Science Labs.” *Work and Occupations* 21 (1): 85–126.

Borgman, Christine L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, Mass.: MIT Press.

Bowker, Geoffrey C. 2005. *Memory Practices in the Sciences*. Cambridge, Mass.: MIT Press.

Bowker, Geoffrey C., Karen Baker, Florence Millerand, and David Ribes. 2010. “Toward Information Infrastructure Studies: Ways of Knowing in a Networked Environment.” In *International Handbook of Internet Research*, edited by Jeremy Hunsinger, Lisbeth Klastrup, and Matthew Allen, 97–117. Springer Netherlands.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, Mass.: MIT Press.

boyd, danah, and Kate Crawford. 2012. “Critical Questions for Big Data.” *Information, Communication & Society* 15 (5): 662–79.

Busch, Lawrence. 2013. *Standards: Recipes for Reality*. Cambridge, Mass.: MIT Press.

Cambrosio, Alberto, and Peter Keating. 1988. “‘Going Monoclonal’: Art, Science, and Magic in the Day-to-Day Use of Hybridoma Technology.” *Social Problems* 35 (3): 244–60.

Campbell-Kelly, Martin. 1990. “Punch-Card Machinery.” In *Computing Before Computers*, edited by William Aspray, 122–155. Ames, IA, USA: Iowa State University Press.

Collins, H. M. 1974. “The TEA Set: Tacit Knowledge and Scientific Networks.” *Science Studies* 4 (2): 165–85.

Converse, Jean M. 2009. *Survey Research in the United States: Roots and Emergence 1890-1960*. New Brunswick, NJ: Transaction Publishers.

Dagiral, Éric, and Ashveen Peerbaye. 2012. “Les mains dans les bases de données : connaître et faire reconnaître le travail invisible.” *Revue d’anthropologie des connaissances* 6 (1): 191–216.

Denis, Jérôme, and Samuel Goëta. 2017. “Rawification and the Careful Generation of Open Government Data.” *Social Studies of Science*, 47 (5): 604-629

Desrosières, Alain. 2010. *La Politique Des Grands Nombres : Histoire de La Raison Statistique*. La Découverte.

Downey, Greg. 2014. “Making Media Work: Time, Space, Identity, and Labor in the Analysis of Information and Communication Infrastructures.” In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, Cambridge, Mass.: MIT Press.

Edwards, Paul, Geoffrey Bowker, Steven Jackson, and Robin Williams. 2009. "Introduction: An Agenda for Infrastructure Studies." *Journal of the Association for Information Systems* 10 (5): 364–74.

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

Edwards, Paul N., Steven Jackson, Melissa Chalmers, Geoffrey C. Bowker, Christine L Borgman, David Ribes, Matt Burton, and Scout Calvert. 2013. "Knowledge Infrastructures: Intellectual Frameworks and Research Challenges." University of Michigan School of Information. http://deepblue.lib.umich.edu/handle/2027.42/97552.

Garfinkel, Harold, Michael Lynch, and Eric Livingstone. 1981. "I.1 The Work of a Discovering Science Construed with Materials from the Optically Discovered Pulsar." *Philosophy of the Social Sciences* 11 (2): 131–58.

Gitelman, Lisa. 2013. *"Raw Data" Is an Oxymoron*. Cambridge, Mass.: MIT Press.

Hey, Tony, Kristin Tolle, and Stewart Tansley. 2009. *The Fourth Paradigm Data-Intensive Scientific Discovery*. Redmond, Wash.: Microsoft Research.

Karasti, Helena, Florence Millerand, Christine M. Hine, and Geoffrey C. Bowker. 2016. "Knowledge Infrastructures: Part I." *Science & Technology Studies* 29 (1): 1–11.

Latour, Bruno, and Stève Woolgar. 1979. *Laboratory Life: The Social Construction of Scientific Facts*. Princeton University Press.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago ; London: University of Chicago Press.

Lynch, Michael E. 1982. "Technical Work and Critical Inquiry: Investigations in a Scientific Laboratory." *Social Studies of Science* 12 (4): 499–533.

Millerand, Florence. 2012. “La science en réseau.” *Revue d’anthropologie des connaissances* 6 (1): 163–90.

Plantin, Jean-Christophe, Carl Lagoze, and Paul N. Edwards. 2018. “Re-Integrating Scholarly Infrastructure: The Ambiguous Role of Data-Sharing Platforms.” *Big Data & Society* 5 (1): 1–14.

Porter, Theodore M. 1986. *The Rise of Statistical Thinking, 1820-1900*. Princeton University Press.

Price, Derek deSolla. 1984. “Of Sealing Wax and Strings.” *Natural History* 1: 49–56.

Ribes, David, and Thomas Finholt. 2009. “The Long Now of Technology Infrastructure: Articulating Tensions in Development.” *Journal of the Association for Information Systems* 10 (5): 375–98.

Shankar, Kalpana, Kristin R. Eschenfelder, and Greg Downey. 2016. “Studying the History of Social Science Data Archives as Knowledge Infrastructure.” *Science & Technology Studies* 29 (2): 62–73.

Shapin, Steven. 1989. “The Invisible Technician.” *American Scientist* 77 (6): 554–63.

Star, Susan. 2002. “Infrastructure and Ethnographic Practice: Working on the Fringes.” *Scandinavian Journal of Information Systems* 14 (2): 107–22.

Star, Susan Leigh, and Karen Ruhleder. 1996. “Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces.” *Information Systems Research* 7: 111–134.

Star, Susan Leigh, and Anselm Strauss. 1999. “Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work.” *Computer Supported Cooperative Work (CSCW)* 8 (1–2): 9–30.

Timmermans, Stefan. 2003. “A Black Technician and Blue Babies.” *Social Studies of Science* 33 (2): 197–229.

Wallis, Jillian C., Elizabeth Rolando, and Christine L. Borgman. 2013. “If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology.” *PLoS ONE* 8 (7): e67332.

Wynholds, Laura A., Jillian C. Wallis, Christine L. Borgman, Ashley Sands, and Sharon Traweek. 2012. “Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices.” In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 19–22. JCDL ’12. New York, NY, USA: ACM.

Table 1. Reconstitution of the "pipeline" for data processing by the author

Action	**1. Deposit the dataset**	**2. Dispatch**	**3. Review and Process**		**4. Contact with the PI (optional)**	**5. Metadata and Formatting**		**6. Verification**	**7. Publication**
Description	The PI or acquisition department deposit a study for processing	The manager review and dispatch the study to a processor	The processor first review the data, identify the problems, and draws a processing plan	The processor then "fixes" the problems: "wild codes," missing values, questions labels, etc.	The processors, after contact with the manager, contacts the PI	The processor writes the metadata for the study	The processor format the datasets and the documents according to templates	The processors send all the files to a manager and another processor for "Quality check"	Once reviewed, the manager approves the publication of the study on the website
Staff	Principal Investigator (PI)	Manager	Processor		PI/Processor/ Manager	Processor		Processor/Manager	Processor
Tool	Deposit form	internal workspace	Scripts, Unix, notepad, SPSS, "eyeball"		Email, spreadsheet	PDF, Hermes, internal workspace		Unix, PDF, notepad, SPSS	internal workspace