BIG DATA & SOCIETY

Original Research Article

Big Data & Society
January–June: 1–12
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517211007510
journals.sagepub.com/home/bds
SAGE

The data archive as factory: Alienation and resistance of data processors

Jean-Christophe Plantin

Abstract

Archival data processing consists of cleaning and formatting data between the moment a dataset is deposited and its publication on the archive's website. In this article, I approach data processing by combining scholarship on invisible labor in knowledge infrastructures with a Marxian framework and show the relevance of considering data processing as factory labor. Using this perspective to analyze ethnographic data collected during a six-month participatory observation at a U.S. data archive, I generate a taxonomy of the forms of alienation that data processing generates, but also the types of resistance that processors develop, across four categories: routine, speed, skill, and meaning. This synthetic approach demonstrates, first, that data processing reproduces typical forms of factory worker's alienation: processors are asked to work along a strict standardized pipeline, at a fast pace, without acquiring substantive skills or having a meaningful involvement in their work. It reveals, second, how data processors resist the alienating nature of this workflow by developing multiple tactics along the same four categories. Seen through this dual lens, data processors are therefore not only invisible workers, but also factory workers who follow and subvert a workflow organized as an assembly line. I conclude by proposing a four-step framework to better value the social contribution of data workers beyond the archive.

Keywords

Data workers, invisible labor, data archive, knowledge infrastructure, data processing, alienation

Introduction

On my first day as a data processor in a U.S. data archive, my newly assigned mentor Sarah gave me a tour of the processing unit.[1] After she introduced me to the other employees and showed me my shared office space, we stopped by the office supply cabinet next to the copy machines. She gave me a quick overview of the items inside, mentioning that this was where to find aspirins. She punctuated her description with a smile, as a friendly yet clear warning of the future headaches to come in this work. After smiling back, I noticed a hammer on the lower shelf of the same cabinet. As I mentioned the incongruity of this tool among pens, notebooks, and colorful sticky notes, she wryly stated, "It is for tough datasets."

It did not take me long to experience what this vignette illustrates: the industrial dimension of working as a data processor. After the tour of the facility, I was promptly put to work and learnt for the next six months what was referred to internally as the "pipeline," which designates the set of standardized procedures to prepare datasets between their reception and their publication. The work of data processors consists of restructuring and formatting data according to specific archival standards, and takes place between the moment a dataset is deposited for archiving, typically by a researcher, and its publication on the institution's website. During this process, my teammates and I assessed the quality of deposited datasets, entered the metadata, restructured them if needed, and formatted the documentation. We executed these repetitive

Department of Media and Communications, London School of Economics and Political Science, London, UK

Corresponding author:
Jean-Christophe Plantin, Department of Media and Communications, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK.
Email: J.plantin1@lse.ac.uk

tasks at a fast pace, without choosing the datasets we worked on, and without acquiring specific archival skills. When one of the managers of the unit described during an interview our job as a “production type of work” organized as an “assembly line,” I knew exactly what he meant. This article takes my former manager’s phrase as an invitation to explore how a factory analytical framing contributes to the theorization of the status of data workers, drawing evidence from the specific case of processors in a data archive.

Researchers in information studies and science and technology studies have amply studied data work in the context of knowledge infrastructures (Borgman, 2010; Edwards et al., 2013; Jackson et al., 2007). This scholarship is organized around the following paradox. On the one hand, it highlights that data workers provide the necessary care, maintenance, and repair (Graham and Thrift, 2007; Jackson, 2014; Mattern, 2014; Russell and Vinsel, 2018) required for the dissemination, archiving, and reuse of data. Downey (2014) showed, for example, how specific “information labor” is required to move data between contexts of production and of reuse. Leonelli (2016) similarly demonstrated how data curators are responsible for “packaging” data in biological research and hence allowing their sharing and reuse by other researchers. The recent analysis of a plurality of data sharing initiatives revealed how dedicated data preparation is primordial to its publication and reuse—either at the scale of private individuals (Pink et al., 2018), design collectives (Baker and Karasti, 2018), local governments (Denis and Goëta, 2017), or communities involved in information management (The Information Maintainers, 2019).

On the other hand, however, the same scholarship has long noted the sheer disconnect between the contribution of data workers to data circulation and the lack of appropriate acknowledgement or reward for their role. As Puig de la Bellacasa put it, care work involves “labours that are often associated with exploitation and domination” (Puig de la Bellacasa, 2012: 198). Ensmenger describes maintenance in software development as “difficult, unpopular, and professionally unrewarding” (Ensmenger, 2014: 8). Historians and sociologists of science have similarly foregrounded how the crucial contribution of laboratory technicians to scientific discovery remains invisible (Barley and Bechky, 1994; Shapin, 1989). When applied to contemporary knowledge infrastructures, this lens reveals how a wide range of data workers—such as information and database managers (Dagiral and Peerbaye, 2012; Millerand, 2012) or staff scanning documents (Chalmers and Edwards, 2017; Ringel and Ribak, 2020)—all remain invisible to virtually anyone outside their workplace.

However, studying data work through this tension has two limitations, which this article aims to remedy. While this scholarship has already shed light on external factors to explain the invisible status of data work—such as the role of race (Timmermans, 2003), gender (Star and Strauss, 1999; Strauss, 1988), or of the hierarchical division of status in science (Shapin, 1989)—it has so far not paid enough attention to the alienating effects of the organization of data work itself. As shown in this paper, by organizing data processing as an assembly line, data archives reproduce forms of worker alienation that have traditionally been associated to Taylorized work. Based on participatory observation in a data archive, this article details how working in this setting can prove as alienating as working on an assembly line. In both settings, work is based on the repetition of meaningless tasks, on pressing time constraints, and on following standardized procedures.

Second, current research on data workers frequently operates by shedding light on otherwise concealed workers and processes, a method often referred to as “infrastructural inversion” (Bowker and Star, 1999). While this is a laudable goal, it can neglect the creativity, skills, and knowledge that processors already employ to resist their work conditions. As I have shown elsewhere (Plantin, 2019), processors have already internalized their position of intermediary workers bound to remain invisible, and scarcely contest it. However, looking in this article at processors as factory workers reveals all the innovative means they deploy in their everyday work, allowing them to resist the fast temporality, the meaninglessness, and the repetitive nature of data processing.

To substantiate this figure of the data archive as a factory, I rely on an autonomist Marxist view on labor in post-industrial societies. This perspective has shown how forms of labor organization typically associated with factory work persist in a post-industrial economy (Tronti, 1962). However, the most recent applications of this theory emphasize the immateriality of contemporary forms of labor (Hardt and Negri, 2000, 2004; Lazzarato, 1996) to the point of overlooking the persistence of Taylorized forms of labor exploitation (Dyer-Witheford, 2001). Using the case of data processors in archive, I show that immaterial labor and factory work can co-exist in the same work setting. I do so by revealing how data processors experience such organization of their labor, and describe it as alternating between alienation and resistance to a Taylorized workflow. To substantiate the former aspect, I use Marx’ definition of estranged labor (Marx and Engels, 1978[1844]) to conceptualize alienation in data processing, with the goal of revealing how, similar to a factory setting, data processors in the archive are made “alien” to the result of their labor. Concerning the latter aspect, the concept of micro-resistance (Scott, 1985) allows me to reveal the multiple tactics that processors develop to counter such alienating labor.

This article is based on ethnographic fieldwork conducted at a U.S. archive that specializes in social science data (Borgman et al., 2018; Eschenfelder and Shankar, 2017; Shankar et al., 2016). I conducted participant observation at the processing unit of this data archive, where I worked as an unpaid part-time intern for six months in 2014. I was granted access to the site by the then director of the institution for a fixed, six-month position, and my dual status as data processor and researcher was made public to all the members of the archive. This participant observation was complemented by 15 semi-structured interviews, conducted in 2014, with the 8 data processors working at the processing unit at the time and with 7 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. I analyzed the data—consisting of verbatim quotes from interviews, notes from participatory observations, manuals and internal documents—through thematic analysis where I focused specifically on how data processing is organized in the data archive, and how it is experienced and felt by data processors.

The article is organized as follows. In the coming section, I present the conceptual framework that sustains my argument and how it applies to the context of the data archive. I start the result section by describing the field site and the data processing pipeline. I then show that looking at the data archive as a factory reveals how the work of data processors oscillates between alienation and resistance, across four categories: routine, speed, skill, and meaning. In the conclusion, I go beyond the data archive and open up a reflection on possible alternative organizations of data work, from the least to the most radical, in order to make the contribution of data workers both visible and socially valued.

Theorizing data processing as factory work

Data processing possesses many of the appearances of a “white-collar” job. A typical work day is from 9 a.m. to 5 p.m. Processors spend most of their time working in front of their computers, interrupted by frequent coffee breaks and team meetings. Describing this activity as factory work requires a deeper look at the labor process, as I do later in this article, along with defining what is meant by factory labor.

Early autonomist Marxists stated the hypothesis of the “social factory” through which capitalism in a post-industrial society radically expands the exploitation of labor from the factory to society as a whole (Tronti, 1962). While manual labor seems to recede in Western economies, they contend that the capitalist exploitation of labor most directly associated with factory work is actually being generalized across society. The paradox that results is the difficulty to see the “specific traits of the factory” as they are now “lost within the generic traits of society” (Tronti, 1962). The task of autonomist Marxists is therefore to reveal the invisible yet widespread forms of factory work (and related exploitation) in societies. A more recent example of such endeavor posits immaterial labor at the center of contemporary economies. Such labor “produces the informational and cultural content of the commodity” and requires dedicated skills “involving cybernetics and computer control” (Lazzarato, 1996: 142). Communication networks generate a form of labor that “creates immaterial products, such as knowledge, information, communication, a relationship, or an emotional response” (Hardt and Negri, 2004: 108). Immaterial labor therefore expands the capitalistic extraction of value in two ways: First, in terms of the commodities that are produced (such as immaterial goods); second, with the type of labor that is commodified (workers sell not only their manual workforce, but also e.g. their creativity, their sense of initiative, or their communicative capacities).

The paradox of autonomist Marxist theories is that by expanding the factory to the society as a whole, and by focusing on contemporary forms of immaterial labor, they lose sight of the persistence of factory-like forms of labor (Dyer-Witheford, 2001). Processing datasets in an archive offers a specific case where the factory remains the central mode of organization of immaterial labor. On the one hand, data processors leverage networked technologies (computers, database, servers) to produce immaterial commodities such as datasets, metadata, documentations, as well as more complex added value, such as the trust that future users have in the deposited datasets, or the prestige that the social science community grants to the archive. On the other hand, data processing is organized by the archive managers and is felt by the data processors as a factory job. Data processors are, in this regard, less similar to the creative workers described by current autonomist thinkers, and very close to the clerical workers described by Bravermann (1988 [1974]). As he put it, once Taylorized principles are applied to clerical work—such as the separation of conception and execution, or the parcelization of labor into repetitive tasks—even this perceived “white-collar” job becomes “just as much a site of manual labor as the factory floor" (p. 218). The same applies to data processing: despite possessing all the aspects of office workers, processors perform a series of routinized tasks along a predefined series of stages that has strong resemblance with working on an assembly line.

Data processors do not passively submit to the Taylorized organization of their labor, though, and find multiple ways to resist such a standardized and repetitive job. The ways data processors counter the alienating nature of their work is not through political struggle or organized solidarity, as the Marxist tradition would emphasize, but by employing multiple forms of everyday resistance. As Scott theorized (1985), class struggle can take the form of "routine resistance" targeting the sources of unequal power while remaining subtle enough to be undetected, e.g. poaching, squatting, deserting, tax evasion, or foot-dragging (Scott, 1985). This focus on micro-resistance reveals the variety of means processors use to contest and mitigate the alienating nature of their work. As the findings will show, these range from hanging out and chatting to more elaborate tactics, such as interchanging the stages of the pipeline, pausing and exploring the datasets, and even developing expertise or pride in their processing.

Alienation and resistance on the data processing pipeline

The stages of data processing in a data archive

The rationale to create archives dedicated to social science data can be traced back to the rise of large-scale quantitative studies starting in the US after the 1940s. Notable examples include the Roper Poll, the General Social Survey, the American National Election Studies, and the Current Population Survey (Converse, 2009). Early initiatives to acquire these datasets were replaced by a wider effort in the 1960s to promote comparative and longitudinal secondary research (Shankar et al., 2016), which led to the establishment of dedicated archives, such as 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 in 1976. Since then, data archives have included a broader range of data, beyond survey, in order to reflect the heterogeneity of data being collected across social sciences. The general mission of data archives—including the one at the center of this article—is to acquire, process, store, and publish datasets collected or deposited to the archive, in order to foster secondary reuse. Data archives can also play a role in professional archival organizations, for example by designing archiving standards and good practices, being involved in research projects, or even providing courses on data analysis. In this context, the specific role of data processors is to format, structure, and prepare datasets for publication in order to foster trust in the datasets during future reuse (Plantin, 2019).

At the time of my fieldwork, the data processing unit consisted of a management team of three members, eight data processors, and other employees who shared their time with other units of the archive (such as technical support). Several interns were also present and learning data processing. These colleagues had almost all recently completed their undergraduate studies, and some were pursuing a part-time Master's degree. They were all hired at the processing unit with an undergraduate degree in social science (mostly psychology and sociology), and some of them had prior experience as research assistants. The majority were women, mostly white and in their early to mid-twenties, and they had been at the archive for at least several months.

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 workstation, and processed datasets. I learned the stages of the data processing pipeline by practicing, reading dedicated internal manuals, attending regular team meetings, and by frequently chatting online with my mentor Sarah. I summarize them as follows:

1. After a researcher (internally simply called "PI" for Principal Investigator) deposits a dataset, or after the archive acquires it and sends it to the processing unit, the data and all the accompanying documents such as codebooks, articles, or other descriptions are reviewed by the management team of the unit, which assesses the level of processing needed and dispatches it to a data processor.
2. After reviewing the dataset—typically survey data—and assessing the amount of work needed to process it, the data processor starts working on the data: this consists of a series of tasks to remove flaws and discrepancies in the structure of the dataset and to format it according to the archive templates. They typically consist of irregularities and formatting issues, such as labels that are missing, incomplete, or that contain irregular characters. The processor's task consists of spotting these discrepancies and reformatting the dataset according to the archive's templates.
3. Once this is completed, all the files are sent to the management team and to a colleague processor for quality control. They each review all the outputs separately to see if all irregularities from the original datasets have been found and fixed, and whether all the documents follow the archive template.

Table 1. Taxonomy of alienation and resistance in data processing.

	Routine	Speed	Skills	Meaning
Alienation	Standardization	Enforcing speed	Deskilling	Meaninglessness
Resistance	Reordering	Socializing	Expertise	Knowledge and pride

4. Once reviewed and approved, the datasets and all accompanying documents are sent for publication and made accessible through the archive website.

Alienation on the assembly line

Working at the archive and processing data, hanging out with my colleagues, or formally interviewing them granted me insights to advance a series of similarities between data processing and factory work, which I describe in this section across four categories (routine, speed, skills, and meaning, cf. Table 1). Using this framework, I first reveal four aspects of the alienating nature of data processing. According to Marx, the division between capital and labor, which characterizes the capitalist economic model, excludes workers from the product of their work. Labor, for Marx, is at the same time an act of production (e.g. of an object) and, in the context of wage labor, is an act of estrangement of workers, through which the result of their labor is something made external to them (Marx and Engels, 1978[1844]).

Similar to the estranged labor described by Marx, processing datasets consists of following highly standardized and routinized procedures, which do not leave room for creativity or expression in the workplace. Data processing consists of working on materials and under a temporality that processors do not have control over, similar to the conveyor belt of the assembly line. Processing data does not result in the acquisition of specific skills, and no matter how long processors keep this job, they do not become specialized in archiving science. Finally, because of all these characteristics, processors question the general meaning and contribution of their work to society. The final dataset processors produce is therefore "something alien" to them (Marx and Engels, 1978[1844]: 71). After presenting this aspect of data processing, the following section will examine four modes of resistance that the same workers employ as a response to tackle each form of alienation (cf. Table 1).

Standardization. Data processing is strictly framed by a series of documents, such as the processing manual, and by routines, which together ensure that all processors work in a similar manner. The tradition of mentorship, during which a new employee will learn processing under the guidance of a senior processor, also reproduces similar ways of working. The final stage of quality control exists to make sure that all the outputs (datasets, documentation, and codebooks) follow the same template. At this stage, if a processor produced a dataset that differs from the expected format, they would be asked to go back to specific stages of the pipeline and to complete it again following established practices. As my fellow data processor Lina told me, what results from this standardization is that "everybody has to do the same kinds of things, exactly. It's a similar process [. . .] and we don't really get specialized at that level."

If this standardization provides a handy series of boxes to tick in order to achieve the desired outcome, it also comes with its downsides. The first is the sense of boredom resulting from a prescribed series of tasks, as described by Lina: "Sitting at your desk for a long period of time is hard. The job can be boring. It can be redundant." The rigidity of the pipeline can also be frustrating for processors who want to innovate. Once, Brett found a way to display more clearly the correspondence between the questions from the codebook and the variables in the dataset. However, because this was not part of the expected formatting, he was asked to redo it according to the guidelines. He expressed his frustration:

> I thought I had found a way to use less time and create a better product [. . .] In the end, I thought the code book would be clearer and have taken me less time to do. And I felt that the way they wanted it to be done would take more time and be equal or worse.

Enforcing speed. Considering data processing as a factory job also reveals the temporality of work at the processing unit. Processors do not have control over the frequency of deposit of datasets at the archive, nor over the rhythm of processing. By design, the archive managers do not know future depositors of studies in advance. By extension, they cannot anticipate when datasets are going to be deposited. The acquisition team of the archive can sometimes proactively acquire datasets, letting the data processing unit know about datasets that are coming. Similarly, some large-scale polls are deposited every year, hence giving an estimated date for future deposit. Beyond these cases, there is no control over the rhythm of submission.

The result is that processors constantly feel pressed for time. As one of the managers of the unit mentioned, "we are trying to get as many [datasets] done as we can every year, because we really report to the council and to the member institutions. And so, we want to get as much out there as possible." This feeling of working against the clock was mainly felt by the management team, as they were the ones receiving incoming datasets and are accountable to the archive's council, but it was also shared among the processors.

Deskilling. As a consequence of the fast pace, there is no time for the processing team to become specialized. The processing unit does not implement specific standards from the archiving world. Rather, they remain at a general processing level. For example, as Steve, a manager, explains:

> I think that [other units of the archive] are able to spend more time and so they develop their specific skills a little bit more. I know [one of them] works with DDI,[2] XML, a lot more than we do. We go through [the pipeline] and when we produce an XML file, we make sure that it's valid, but we don't go in and adjust it. We don't develop that.

Due to the fast pace under which they have to process datasets, there is no possibility for processors to develop skills in the latest archiving standards, something that other specialized units in the archive are able to do. For the same reason, processors do not learn about research methods either. As Lina, a data processor, puts it:

> I just think it would take you a long time to really learn a lot about research method with just your processing [...] because you wouldn't run into these issues, and along the way I think you would need to be doing professional development, learning all these packages. If you are not doing that I don't see anyone really learning a lot about research methods or how to analyze data.

Many processors mentioned how they developed their functional knowledge of the various statistical packages used in the archive. However, because they work only on the structural level of data, they do not develop analytical skills.

Meaninglessness. The combination of the routine nature of the job, lack of transferable skills, and absence of career progression results in processors thinking of their job as temporary. As Steve states:

> Right now, there's not so many people but we have had quite a few that have stayed on for a couple of years, or a year or two. It's only been recently with the downturn of the economy where people have stayed longer because this is not really... [...] It was never thought of as a long-term kind of a career place.

The reduced number of managers needed, most of whom were appointed recently, similarly reduces internal career progression. In addition to these poor career prospects, several processors mentioned that they wanted to find a more meaningful job. This is how Shannon considers her job as data processor:

> No offense, I feel like it doesn't really add value to the world. Of course, it helps professors, to put their data [out there] but in terms of direct help to people? [...] I can see maybe someone getting some data that we process here and writing a paper and maybe someone on Capitol Hill looking at it, but the odds of that are very, very, very low.

Most processors indeed consider this job as transitory toward something else, such as applying to graduate school. The meaninglessness of the job also comes from the division of labor between PI and data processors. The current model of data archiving emphasizes the responsibility of the researchers in preparing their data before submission to an archive. This consists of PIs providing details on the data collection methods, on the structure of data files, or on the control procedures adopted for the deposited dataset. These details are crucial to the data archive, and a dataset well prepared will result in an easier processing and faster publication by the archive. Data archives try to incentivize and guide researchers to prepare their data before submission, for example, by publishing detailed guidelines for data preparation.[3] However, despite these documents, studies on data sharing show that extra costs, time, and labor constitute major disincentives for researchers to adequately prepare their data for sharing (Borgman, 2012; Tenopir et al., 2011).

By contrasting such necessary data preparation before submission and the lack of incentives for depositors to do it, it appears that the role of processors is to compensate for the lack of adequate data preparation by researchers depositing their datasets. This task is a common source of frustration for processors. As Sarah put it: "sometimes we get so bogged down with adding things that [depositors] should have added themselves." While formatting the data for publication requires some knowledge about the institutional templates that only the data processors can possess, the stage of preparing datasets frequently comes with the uneasy feeling of "cleaning up" after someone—as there are many flaws that could have been identified easily and repaired by the depositors themselves before submission—adding to the meaninglessness of the job.

Micro-resistance on the assembly line

Literature on the invisible labor in knowledge infrastructure tends to concentrate on shedding light and giving voice to invisible care workers, typically through an "infrastructural inversion" (Bowker and Star, 1999) taking the form of scholarship, art projects,[4] and manifestos (The Information Maintainers, 2019). However, considering data processing as Taylorized work goes beyond the opposition between visibility and invisibility, and sheds light instead on the multiple ways processors already resist their work conditions. Scholarship in anthropology has complemented the emphasis in Marxian analysis on collective organization of labor to show how class struggle also takes place at the micro-level e.g. in the forms of tactics (Certeau, 1990) or bricolage (Lévi-Strauss, 1962). Similarly, Scott's concept of "micro-resistance" (1985) brings to the fore practices that can appear as mundane and prosaic, but in the context of the workplace, can constitute forms of resistance.

In the archive, data processors similarly developed four types of micro-resistance (presented below and summarized in Table 1): they find ways to re-order the stages that constitute the pipeline, hereby going against its linear and repetitive nature; they take time to socialize, as a way of reclaiming time over a fast-paced job; they develop expertise in a pipeline that rewards fast execution of a series of pre-defined stages; finally, they develop knowledge and pride in their work, for example by exploring datasets instead of working only on their structure.

Reordering. The pipeline is designed in a linear way and requires processors to complete specific tasks before being able to move on to the next ones. Within these stages, however, processors still have the possibility to organize their workflow in different ways. For example, some processors postpone some stages, such as metadata, until they can no longer avoid them. As Shannon puts it: "you can kind of tweak [the pipeline] as you need [...]: I can fill out the basic parts of the metadata and come back, but when you're doing your first quality check, you need to have all of your metadata done." Similarly, when identifying and fixing flaws in datasets, processors like Sarah worked on "designated missing values first or [...] value labels first," with the same end result.

Reordering tasks in the pipeline allows processors to reorganize their workflow in a way that better reflects their personal preferences. For some, it is a way of postponing a less appreciated part of the job. Shannon jokes about her relationship with metadata: "We just don't get along very well," while this stage makes Lina feels she is "in school again" because adding metadata requires a lot of editing and multiple checks. For others, it allows them to choose their preferred technique to execute a specific stage, even if it can be different from the one taught in the processing manuals. For example, Brett prefers using macros instead of regular expressions to structure a dataset, as the latter is not "where [*his*] skill set is." Whatever the reasons, the management team tolerates such departing from the order of processing stages, as long as the final outputs follow the standard templates of the archive.

Socializing. During our shifts, my colleagues and I would take long and frequent breaks. We congregated in the section of the hallway where most of the processors' offices were situated, conveniently located at the other end of the office of the managers. Chats revolved around everyday life, living in the city, sharing gossip about other employees, and sometimes talking about the processing work itself. Nothing is out of the ordinary in this practice. However, what came out of individual interviews was that processors felt even more inclined to engage in this social aspect of the job, as it constituted a way to break the 'boring' routine inherent in processing. Without a break, "it's just me and my data every day," as Lily put it. For some, it even constituted one of the main perks of the job and directly contrasted with their past employment histories. As another processor named Kenneth put it:

> I like being around other people who think the same way when it comes to research and stuff like that. Whereas if I were to go back to driving a truck, I'm not gonna get into too many research conversations.

These informal moments were also opportunities to talk about work, especially when someone is stuck on a stage of the pipeline. For example, Brett likes "trying to troubleshoot with the other employees" through informal chatting. In this workplace where employees are tasked with processing datasets as fast as they can, taking frequent breaks is a way to regain control over the labor discipline shaped by the "clock time" of the factory (Thompson, 1967).

Expertise. While processing is based on an existing pipeline with a series of predetermined tasks, it also demands problem-solving capacities that allow some processors to develop a form of expertise and even a reputation in the unit. Some datasets require extensive reformatting or come with flaws that are not directly identifiable. Kenneth mentioned how he especially likes the challenge "when there's a study that's messed up and [he has] to try to piece it together." He is not sure why he is specifically good with these datasets—he mentioned a vague capacity to see "the big picture." Still, he has gained over time a reputation among his peers and managers for being specialized in such challenging datasets—that he calls "disaster studies"—to the point of being a running joke with his manager. He remembers him facetiously saying once, "I'm sure I'll have some disaster, clunky studies to give you."

The organization of the pipeline largely precludes personal initiatives, making all processors in theory interchangeable with each other. The fact that some processors possess an expertise that makes them actually stand out, and that they claim this reputation in the archive, is a form of resistance against an anonymizing organization of labor that levels each processor's contribution.

Knowledge and pride. By design, data processing concerns only the structure of the data and does not require in-depth analysis of the content, methodology, or results of a study. However, processors have to develop some minimal knowledge about the formal description of the study—typically the author, type of data, and the context of collection. This information is gathered by looking at the various documents coming with the study (such as published papers) and sometimes from searching on the internet. This knowledge is necessary and part of the job, as processors need eventually to write the metadata and other descriptions of the study for the archive's website. As Sarah told me early on during my internship, "when you work on a study, you're the person in the archive who knows that study the best." This knowledge remains, however, ad hoc and temporary, and will last only as long as the processor is working on the study.

Processors also develop another somewhat unexpected type of resistance that comes from direct exploration of datasets in their spare time and not as part of their processing. It either consists of looking at the frequencies of the datasets they are processing, or doing basic analyses such as cross-tabulating two variables. In this way, Kenneth could, for instance, learn about the differences between the international perception of the U.S. President and U.S. foreign policy:

> I would think that people would paint it with a broad brush. If you don't like American foreign policy, then you don't like American politicians, you don't like Barack Obama and you don't like the American public [. . .] but it's actually, surprisingly, not like that.

Lily similarly spent time exploring a dataset on dating, in which:

> they had asked respondents: 'what would be the most convincing pick-up line?' [She laughs] And I don't remember the responses but, it was funny, and I remember thumbing through and wanting to work on that project.

This type of data exploration is not part of the pipeline and is not directly useful for processing, but still constitutes a way of compensating for the distant view on data and the boredom that processing involves. While the pipeline restricts the work of processors at the structural level of data, taking time to explore and learn from the actual content of datasets acts as a form of micro-resistance.

Finally, when the study has gone through all the stages of the pipeline and passed the final quality control, processors experience a feeling of relief that they do not have to work on that study anymore—but also a sense of pride in delivering a product whose quality is certified by high-quality standards. As Caroline put it:

> It's just kind of satisfying to get [the dataset] to the point where you've checked all your errors. You have no warnings, no errors. Everything looks good. You've designated all your missing [values] and [it's] just like having the complete finished product.

This pride in a well-executed job echoes the sense of altruism, mentioned earlier, coming from making social science better through creating validated datasets. Brett told me how "cracking data mistakes [is] important to [him] because that's the actual data people are gonna be working with."

Many have warned against a tendency to romanticize the heroic actions of social actors through the concept of micro-resistance (Abu-Lughod, 1990; Gal, 1995). Moreover, as autonomist Marxists have pointed out, the extension of the factory to the society as a whole means that even resistance already falls within the reach of capitalism. Similarly, the forms of resistance in the data archive are not felt nor presented as heroic, and they *in fine* do not change anything to the way the pipeline is organized. However, if resistance is less relevant as a means to actually change things, it offers a powerful diagnostic of power (Abu-Lughod, 1990). In the case of the archive, the study of micro-resistance shows how power relations are organized in the data archive, resulting at the same time in an alienating organization of data processing that processors contest by carving out time and space to find meaning in their work.

Conclusion and suggestions to value data workers' contributions

The analysis of data processing in an archive complements research that has demonstrated the importance of care, repair, and maintenance applied to data work (Baker and Karasti, 2018; Denis and Goëta, 2017; Pink et al., 2018; The Information Maintainers, 2019). In the context of knowledge infrastructure, such research typically focuses on uncovering the contribution of invisible labor to knowledge production (Downey, 2002; Star and Strauss, 1999; Strauss, 1988; Suchman, 2007; Timmermans, 2003). This article contributes to these two bodies of research by showing that the low status of data processors does not result only from factors such as race, gender, and hierarchical division in academia, but also from the very organization of labor. Relying on Autonomists' theories on the society as factory and on immaterial labor (Dyer-Witheford, 2001; Lazzarato, 1996; Michael and Negri, 2004; Tronti, 1962) to analyze results from the ethnography of a data archive yields two main results. First, the organization of data processing in an archive results in a workflow very similar to an assembly line. Data processors are not only invisible and unacknowledged data workers, but also factory workers experiencing forms of alienation typical of Taylorized work, across the four dimensions of routine, speed, skills, and meaning. Second, I show that the same alienating procedures generate forms of micro-resistance (Scott, 1985). Data processors develop creative tactics to resist the boredom and meaninglessness of their job, ranging from taking frequent breaks to developing specific knowledge around data processing, and even taking pride in the result of their labor. Data archives can be seen simultaneously as an institution allowing the circulation of data and knowledge, but also as a factory relying on low status and repetitive labor to deliver reliable datasets. In this context, data processors are alienated twice: their capacities for initiative and creativity when processing datasets in the archive are prohibited by a strict and standardized framework; and their larger contribution to social science remains equally unacknowledged, despite the essential work they provide to create trust in datasets.

Compared to contemporary forms of "digital Taylorism" such as micro-work (Altenried, 2020; Gray and Siddharth, 2019) or "gig work" in the platform economy (Huws, 2019; Woodcock and Graham, 2019), everyday work at the data archive is relatively protected. The archive where I worked is related to the local university, and therefore provides benefits and relative employment security to its employees—a strong motivation for processors to apply and stay on the job. Workplace surveillance is almost non-existent, beyond the pressure to process datasets on time. The secure environment of the data archive, and the types of micro-resistance that are developed there, provide a productive setting to experiment with new ways of bringing visibility and acknowledgement to data workers beyond the data archive—for example in the forms of Mechanical Turk workers (Irani, 2015), content moderators (Gillespie, 2018; Roberts, 2019), or micro-workers training artificial intelligence algorithms (Tubaro et al., 2020). Moreover, using the factory as a heuristics to study labor politics of data work takes part in a wider effort in critical library and information studies to bring together archiving with social theory (Waterton, 2010), materiality (Stuchel, 2020) and feminist epistemologies (Caswell, 2020).

Answering this call to embrace the normative dimension of scholarship, results from this ethnography offer a four-stage roadmap (cf. Table 2) to acknowledge and value data work and imagine what an emancipatory organization of this work could be. These stages range from the least to the most radical solutions: making data workers visible, acknowledging their contribution, building collective organization, and contesting the division of labor between data producers and data workers.

Making data workers visible

The first step consists of uncovering the presence of invisible data work. This can be achieved through scholarly publications—as witnessed by the existing scholarship cited earlier on invisible labor in knowledge infrastructures—or through other means such as investigative journalism, documentary films, or art projects. Beyond the archive, other contemporary cases include pseudo-AI (Tubaro et al., 2020), where a service presented as automated relies on hidden data entry workers. This step is the most feasible and least radical option. However, its reach can vary (ranging from a mostly academic audience for scholarly publication to a hopefully wider public with art projects), and this visibility originates from outside actors, not from the concerned workers themselves.

Acknowledging their contribution

The second step is more ambitious and consists of showing not only that data workers exist, but also how their contribution is essential to the functioning of the socio-technical systems of which they are part (e.g. social media platforms). At the level of the archive, it could consist of adding the names of the processors involved in the published dataset (currently not mentioned), or of a dedicated section of the archive's website presenting who the processors are

Table 2. Four-step framework for the emancipation of data workers.

	Goals	Means	Key scholarship	Central figure
Making data workers visible	Showing the existence of data workers	Scholarship, art, investigative journalism	Knowledge infrastructure studies	Invisible technician
Acknowledging their contribution	Understanding their social contribution	Adding names of processors in final product, institutional profile webpage	Social studies of repair and maintenance	Care worker
Building collective organizations	Developing marketable skills and career development	Professional organizations and representatives	Marxian approaches to labor organization	Unionized worker
Contesting the division of labor	Increasing control of data workers over production	Rejecting insufficiently prepared datasets, extending the participation to data preparation	Autonomists, Digital labor theorists	Member of a worker-owned cooperative

and what they do. Some of these initiatives have already been implemented by archives, such as video clips describing curation on the ICPSR's website.[5]

Building collective organizations

The third step moves closer to a radical change of the organization of data work. Building upon a Marxist tradition of labor movement, it would consist of developing professional organizations representing the interests of data workers, similar to a union. In the archive, such organization would aim to improve work conditions (such as allowing contributions to the pipeline), to recognize the skills they develop and those that they should acquire to be competitive for subsequent career mobility. In the archive, that could take the form of offering additional training to processors in the latest archival standards, or developing pathways to further studies at the university. Several initiatives in the "gig economy" are close to this third stage and aim to adapt traditional forms of workers' representation to platform-mediated labor (Graham and Shaw, 2017).

Contesting the division of labor

The last and most radical step consists of data workers gaining deeper control over the data processing pipeline. While it is yet to see what a collective ownership of the means of production would be in this case (with processors owning the data archive?), current experiments on platform cooperativism (Scholz and Schneider, 2017) provide the inspiration to contest the current division of labor between the data producers and data processors. As datasets insufficiently prepared before submission generate additional menial work for processors, one could imagine processors sending those datasets back to the depositor for extra preparation before resubmission. Data archives could also link the deposit of datasets to the participation of data producers in the processing of one or parts of a dataset, just like a food cooperative periodically asks its members to volunteer to stacking goods.

These are four speculative and provocative steps that can hopefully trigger reflections and actions, either at the level of the archive—among archive and repository managers, information scientists, critical library and information studies researchers, and data processors themselves—or with other types of data work to potentially mitigate its alienating nature and hopefully to value adequately its contribution to society.

Acknowledgements

I would like to thank Meghanne Barker, Bart Cammaerts, Nick Couldry, Olivier Driessens, Paul N. Edwards, James R. Jacobs, Carl Lagoze, Robin Mansell, Christian Sandvig, Andrew Schrock (Indelible Voice), and ShinJoung Yeo for their useful comments on earlier versions of this research, as well as the *Big Data & Society* editors and anonymous reviewers for the quality of their feedback.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article results from research conducted during the following project: New directions in the Study of Infrastructures (2013–2014, Principal Investigators: Prof Paul Edwards, Dr Carl Lagoze, Prof Christian Sandvig, funded by the MCubed program, University of Michigan).

ORCID iD

Jean-Christophe Plantin https://orcid.org/0000-0001-8041-6679

Notes

1. All the names have been changed.
2. “The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences,” cf. DDI Alliance website: https://ddialliance.org/
3. For example, the UK Data Archive: *Managing and Sharing Research Data: Best Practices For Researchers*; Irish Social Science Data Archive: *ISSDA Services for Depositors & Researchers*; ICPSR: *Guide to Social Science Data Preparation and Archiving*.
4. Such as the Project *Mama* that aims to uncover everyday maintenance through “storytelling, including research, drawings, writings, workshops, conferences, exhibitions and direct action.” Cf.: http://mama.brussels/office.html
5. Data Management & Curation, ICPSR: https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/index.html

References

Abu-Lughod L (1990) The romance of resistance: Tracing transformations of power through Bedouin women. *American Ethnologist* 17(1): 41–55.

Altenried M (2020) The platform as factory: Crowdwork and the hidden labour behind artificial intelligence. *Capital & Class* 44(2): 145–158.

Baker KS and Karasti H (2018) Data care and its politics: Designing for local collective data management as a neglected thing. In: *Proceedings of the 15th participatory design conference: Full papers – Volume 1*, New York, NY, USA, 2018, pp.10:1–10:12. New York: ACM.

Barley S and Bechky B (1994) In the backrooms of science: The work of technicians in science labs. *Work and Occupations* 21(1): 85–126.

Borgman CL (2010) *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge: The MIT Press.

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

Borgman CL, Scharnhorst A and Golshan MS (2018) Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *arXiv:1802.02689 [cs]*. Available at: http://arxiv.org/abs/1802.02689

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

Braverman H (1988) *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. New York: Monthly Review Press.

Caswell M (2020) Dusting for fingerprints: Introducing feminist standpoint appraisal. *Journal of Critical Library and Information Studies* 3: (pre-prints).

Certeau MD (1990) *L'invention Du Quotidien, Tome 1 : Arts de Faire*. Paris: Gallimard.

Chalmers MK and Edwards PN (2017) Producing “one vast index”: Google Book Search as an algorithmic system. *Big Data & Society* 4(2): 205395171771695.

Converse JM (2009) *Survey Research in the United States: Roots and Emergence 1890–1960*. New Brunswick: Transaction Publishers.

Dagiral É and Peerbaye A (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 J and Goëta S (2017) Rawification and the careful generation of open government data. *Social Studies of Science* 47(5): 604–629.

Downey G (2014) Making media work: Time, space, identity, and labor in the analysis of information and communication infrastructures. In: Gillespie T, Boczkowski PJ and Foot KA (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge: The MIT Press, pp.141–165.

Downey GJ (2002) *Telegraph Messenger Boys: Labor, Technology, and Geography, 1850–1950*. 1st ed. New York: Routledge.

Dyer-Witheford N (2001) Empire, immaterial labor, the new combinations, and the global worker. *Rethinking Marxism* 13(3–4): 70–80.

Edwards PN, Jackson S, Chalmers M, et al. (2013) *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. University of Michigan School of Information. Available at: http://deepblue.lib.umich.edu/handle/2027.42/97552 (accessed 30 March 2021)

Ensmenger N (2014) When good software goes bad. The surprising durability of an ephemeral technology. In: *MICE (mistakes, ignorance, contingency, and error) conference*, Munich, Germany, 2–4 October 2014. Available at: http://homes.soic.indiana.edu/nensmeng/files/ensmenger-mice.pdf

Eschenfelder K and Shankar K (2017) Organizational resilience in data archives: Three case studies in social science data archives. *Data Science Journal* 16: 12.

Gal S (1995) Language and the ‘Arts of Resistance’. *Cultural Anthropology* 10(3): 407–424.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Graham M and Shaw J (eds) (2017) *Towards a Fairer Gig Economy*. London: Meatspace Press. Available at: https://meatspacepress.com/towards-a-fairer-gig-economy/

Graham S and Thrift N (2007) Out of order: Understanding repair and maintenance. *Theory, Culture & Society* 24(3): 1–25.

Gray ML and Siddharth S (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Hardt M and Negri A (2000) *Empire. Cloth/Dust Jacket and Mylar Wrapped Octavo Edition*. Cambridge: Harvard University Press.

Huws U (2019) *Labour in Contemporary Capitalism: What Next? Dynamics of Virtual Work*. London: Palgrave Macmillan.

Irani L (2015) Justice for "Data Janitors". Available at: http://www.publicbooks.org/nonfiction/justice-for-data-janitors (accessed 19 August 2016).

Jackson S (2014) Rethinking repair. In: Gillespie T, Boczkowski PJ and Foot KA (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge: The MIT Press, pp. 221–240.

Jackson SJ, Edwards PN, Bowker GC, et al. (2007) Understanding infrastructure: History, heuristics and cyberinfrastructure policy. *First Monday* 12(6).

Lazzarato M (1996) Immaterial labor. In: Virno P and Hardt M (eds) *Radical Thought in Italy : A Potential Politics*. Minneapolis: University of Minnesota Press, pp. 142–157.

Leonelli S (2016) *Data-Centric Biology: A Philosophical Study*. Chicago; London: University of Chicago Press.

Lévi-Strauss C (1962) *La Pensée Sauvage*. Germany: Plon.

Marx K and Engels F (1978) *The Marx-Engels Reader (ed. RC Tucker)*. New York: W. W. Norton & Company.

Mattern S (2014) Library as infrastructure. *Places Journal*. Available at: https://placesjournal.org/article/library-as-infrastructure/ (accessed 30 March 2021).

Michael H and Negri A (2004) *Multitude: War and Democracy in the Age of Empire*. Annotated edition. New York, NY: Penguin Group.

Millerand F (2012) La science en réseau. *Revue d'anthropologie des connaissances* 6(1): 163–190.

Pink S, Ruckenstein M, Willim R, et al. (2018) Broken data: Conceptualising data in an emerging world. *Big Data & Society* 5(1).

Plantin J-C (2019) Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. *Science, Technology, & Human Values* 44(1): 52–73.

Puig de la Bellacasa M (2012) 'Nothing comes without its world': Thinking with care. *The Sociological Review* 60(2): 197–216.

Ringel S and Ribak R (2020) 'Place a book and walk away': Archival digitization as a socio-technical practice. *Information, Communication & Society*. Epub ahead of print 22 May 2020. https://doi-org.gate3.library.lse.ac.uk/10.1080/1369118X.2020.1766534 (accessed 30 March 2021).

Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.

Russell A and Vinsel L (2018) After innovation, turn to maintenance. *Technology and Culture* 59(1): 1–25.

Scholz T and Schneider N (eds) (2017) *Ours to Hack and to Own: The Rise of Platform Cooperativism, a New Vision for the Future of Work and a Fairer Internet*. New York: OR Books.

Scott JC (1985) *Weapons of the Weak: Everyday Forms of Peasant Resistance*. New Haven: Yale University Press.

Shankar K, Eschenfelder KR and Downey G (2016) Studying the history of social science data archives as knowledge infrastructure. *Science & Technology Studies* 29(2): 62–73.

Shapin S (1989) The invisible technician. *American Scientist* 77(6): 554–563.

Star SL and Strauss A (1999) Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer Supported Cooperative Work (CSCW)* 8(1–2): 9–30.

Strauss A (1988) The articulation of project work: An organizational process. *The Sociological Quarterly* 29(2): 163–178.

Stuchel D (2020) Material provocations in the archives. *Journal of Critical Library and Information Studies* 3(1): (pre-prints).

Suchman LA (2007) *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge; New York: Cambridge University Press.

Tenopir C, Allard S, Douglass K, et al. (2011) Data sharing by scientists: Practices and perceptions. *Plos ONE* 6(6): e21101.

The Information Maintainers (2019) Information maintenance as a practice of care. Available at: https://doi.org/10.5281/zenodo.3236410 (accessed 30 March 2021).

Thompson EP (1967) Time, work-discipline, and industrial capitalism. *Past and Present* (38): 56–97.

Timmermans S (2003) A black technician and blue babies. *Social Studies of Science* 33(2): 197–229.

Tronti M (1962) Factory and society. In: *Operaismo in English*. Available at: https://operaismoinenglish.wordpress.com/2013/06/13/factory-and-society/ (accessed 30 March 2021).

Tubaro P, Casilli AA and Coville M (2020) The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society* 7(1).

Waterton C (2010) Experimenting with the archive: STS-ers as analysts and co-constructors of databases and other archival forms: *Science, Technology, & Human Values* 35(5): 645–676.

Woodcock J and Graham M (2019) *The Gig Economy: A Critical Introduction*. 1st ed. Cambridge; Medford: Polity.