Surfacing Data Change in Scientific Work

Drew Paine(✉) and Lavanya Ramakrishnan

Data Science and Technology Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
{pained,lramakrishnan}@lbl.gov

Abstract. Data are essential products of scientific work that move among and through research infrastructures over time. Data constantly changes due to evolving practices and knowledge, requiring improvisational work by scientists to determine the effects on analyses. Today for end users of datasets much of the information about changes, and the processes leading to them, is invisible—embedded elsewhere in the work of a collaboration. Simultaneously scientists use increasing quantities of data, making ad hoc approaches to identifying change difficult to scale effectively. Our research investigates data change by examining how scientists make sense of change in datasets being created and sustained by the collaborative infrastructures they engage with. We examine two forms of change, before examining how trust and project rhythms influence a scientist's notion that the newest available data are the best. We explore the opportunity to design tools and practices to support user examinations of data change and surface key provenance information embedded in research infrastructures.

Keywords: Data change · Invisible work · Research infrastructures

1 Introduction

Research infrastructures are long-lasting networks of people, institutions, and artifacts that produce, share, and sustain information about the world [4]. Those enacted for collaborative science are rooted in the data they produce and sustain over time [2,5,10,15]. Data are an essential element of scientific practices that depend on context and individual's interpretations, providing 'monopoly rents' [1] and serving as the 'lifeblood' [10] of this enterprise. Studies of data and research infrastructures highlight the contextually dependent work to produce, process, share, support, and facilitate use and reuse of data over time [1,2,19,25]. This process comes with some amount of friction [4,6] as data moves among different sets of stakeholders, systems, and practices such that the narratives shaping it shift and evolve, making data change a fundamental element of scientific work. Frustratingly these narratives are often invisible when the work and decisions leading to data changes are embedded inside infrastructural processes that end users do not or cannot see.

N. G. Taylor et al. (Eds.): iConference 2019, LNCS 11420, pp. 15–26, 2019.
https://doi.org/10.1007/978-3-030-15742-5_2

Star and Strauss [21] stress that work is not inherently visible or invisible, it depends on the perspective of the person. Consequently, the practices of researchers producing and analyzing data in research infrastructures make (in)visible the change in the products being produced, in part because there are not clearly established guidelines within or across communities for surfacing and sharing this information. Each discipline or infrastructure "has its own norms and standards for the imagination of data, just as every field has its accepted methodologies and its evolved structures of practice" [9]. Some scientific collaborations provide high-level information about changes between versions of datasets while others provide none or minimal. As a result, the work of researchers calculating and sharing data change information is currently often ad-hoc, with inconsistent tools and practices that are often invisible to collaborators and inefficient at large scales. Our study's contribution is to address this gap by examining conceptualizations of data change in research infrastructures so that we might move towards a more systematic set of practices and tools for use within and among infrastructures. This is a formidable design opportunity as the quantities of data scale up and ad-hoc approaches no longer suffice, leading us to ask: *How do scientists make sense of data change in their research work?*

In the remainder of this paper, we discuss work on research infrastructures and scientific data that ground our study, including research on data cleaning as well as invisible work, before describing our research site and methods. Our findings examine some ways participants think about data change and how it shapes their work within infrastructures today. We conclude by discussing opportunities to make this invisible work more visible.

2 Literature Review

Investigations of scientific collaboration and the development and emergence of research infrastructures are commonly theorized using Star and Ruhleder's [20] relational infrastructure lens. This theoretical lens articulates eight facets, from having reach across sites and being learned as part of membership in groups and transparent in use, to existing within other structures through inherent embeddedness. Embeddedness is of particular concern here as we work to surface data change as an invisible, sunk in aspect of everyday scientific work for end users of datasets. To conceptualize data change we need to first articulate what data means in scientific work then explore ways of investigating infrastructure so that we can connect this notion to work on data processing and cleaning.

Data do not arise from nothing. Gitelman [9] emphasizes that "data need to be imagined *as* data to exist and function as such" while Kitchin [12] posits that data are the "material produced by abstracting the world into categories, measures and other representational forms that constitute the building blocks from which information and knowledge are created." In collaborative science, data help define boundaries among stakeholders with different communities of practice, can act as a gateway into different communities, and often indicate status [1]. Scientists must iteratively seek information and narrate their evolving

products to successfully work with data [3,25], and essential to the use and reuse of data is scientist's trust in those who created it and their ability to find relevant information to answer questions about its production [7,19].

Studying infrastructures as a relational process requires a researcher constantly make decisions about what to include and exclude from an inquiry. The researcher, their subjects, and the research context co-construct what is visible and invisible in the study. Karasti and Blomberg [11] emphasize the need to 'construct the field' when undertaking ethnographic inquiries of infrastructures. A key part of constructing a field can be examining what work is visible and invisible. Star and Strauss [21] state that "what exactly counts as work varies a lot" depending on the context and who is viewing the activity at hand. Traditional 'women's work' of taking care of a household was invisible to many classifications, marginalizing this important effort and leaving it out of potential conversations in the design of systems and policies. Star and Strauss emphasize disembedding background work, examining that which is right in front of the observer but not always focused on, to make invisible work visible.

In our study we constructed our field to start surfacing the invisible work behind data change in some infrastructures of scientific research by drawing upon data processing or cleaning studies. Previous work stresses the labor intensive work of data processing or cleaning [14,16–18]. Rawson and Muñoz [18] note that specifics of data cleaning often "reside in the general professional practices, materials, personal histories, and tools of the researchers" rather than explicitly captured and included with a data release. Plantin [16] similarly highlights that cleaning is often invisible work. Earlier, Paine et al. [14] foreground and unpack the intricate, challenging data processing work scientists undertake to clean data by removing or fixing spurious values, selecting subsets of data for particular analyses, and transforming between formats to produce a product that meets their needs. From this body of work we see a gap where information underlying changes to datasets during cleaning and processing may be embedded and invisible to end users, something our work aims to tackle.

3 Research Sites and Methods

Our study[1] is investigating data change in different disciplines at Lawrence Berkeley National Laboratory, a US Department of Energy national lab. A long-standing defining feature of US national labs is a collaborative, often multidisciplinary, approach to research such that all of our subjects participate in projects with members distributed around the US and world. This paper's findings emerge from interviews with subjects in astronomy and earth sciences.

We conducted semi-structured interviews with five astronomers and five earth scientists between October 2017 and February 2018. Interviewees fulfill two general roles (sometimes both): Data Producers, individuals working on producing

[1] This work is part of the Deduce project (http://deduce.lbl.gov). The goal of the Deduce project is to develop methods and tools that support data change exploration and management in the context of data analysis pipelines.

data releases; and, Data Users utilizing data releases for analyses. Astronomy interviews included members of the Sloan Digital Sky Survey (SDSS) and/or the Dark Energy Spectroscopic Instrument (DESI) projects[2] collecting observational data. The earth science projects[3] produce sensor-based observation data at field sites. Some users augment field data with additional satellite data. Our interviewees included four astronomy data producers, one astronomy data user, two earth science data producers/users, and three earth science data users. The SDSS and Ameriflux projects provide high-level information about changes between data releases, but low-level details that would help end users assess potential effects to their analyses are not available with the data at this time.

Interviews were recorded, transcribed, and cleaned by the first author for analysis. They ranged from 58 to 80 min (avg 63 min). Our interview protocol was designed to learn about various aspects of an individual's research projects, focusing on data they work with and how it is obtained. We asked how subjects determine which version of data products to use for an analysis, and effects (both expected and unexpected) from changes in data products. We analyzed our data using a modified grounded theory process, open coding transcripts for responses to our questions and emergent ideas [26] and assessing these codes in relation to the literature identified earlier. Coding enabled us to distinguish sources of data change among our interviewees. We identified common themes such as the general categories of data change for our subjects as well as thoughts on the process of selecting data for use that inform our findings.

4 Findings

Our findings explore multiple facets to data change in scientific infrastructures. We examine interviewee concerns and characterize two types of data change in their work. We then unpack their expectations that newer datasets are better than older by considering our subject's trust in their collaborators and project data release processes. Finally, we examine how a project's rhythms and organizational structures for data release are part of scientist's trust in processes.

4.1 Why Scientists Are Concerned About Data Change

Change in datasets is an expected facet of work for our scientific subjects. The processes leading to different data releases are not always visible to different stakeholders, making it hard to evaluate the effects of nuanced changes. A lack of actionable information makes it hard for these producers and users to assess when they need to re-run a past analysis or adapt a new one as a result of changes to some part of dataset.

Our subjects work with datasets that are continuously expanding, adding new data, even as ongoing data processing work employs different cleaning practices to refine existing data. Data producers putting out new releases need to

[2] http://www.sdss.org/, http://www.desi.lbl.gov/.
[3] http://watershed.lbl.gov/, http://ameriflux.lbl.gov.

be able to check copies mirrored across archives for unexpected changes and evaluate the impact of processing on data values. Science users need to be able to assess whether changed data values will impact their current or past analyses in a significant way. Information about data change also provides necessary provenance information about the data. The lineage or history of the data is critical to allow scientists to make important decisions when processing data. One example comes from an astronomer who leads their project's data release team. They emphasized that while developing a new yearly release the collaboration will reprocess all of the data from the start for big and small changes.

> *"So every year they release an updated set. ... We reprocess all of the data from the start. ... Some years it's just sort of a slight incremental change. You know, fixing one thing here or there and it's just like, for completeness, you rerun it on everything. Other times, it's a fairly major update."* (Astronomy data release manager)

These changes can be due to an incremental or major revision to their scientific approach expressed through various software pipelines. In other cases the collaboration, or a sub-group, shifts their scientific focus. At one point these astronomers began to try to image faint objects rather than bright objects, altering the characteristics of the signals sought and the scientific approach to processing data. This type of shift upends the assumptions they have embedded into their practices and artifacts. The data produced is different as a result of foregrounding issues with their software pipeline that were previously invisible to, or intentionally ignored by, these data producers (and as a result the end users of data) in the course of their work.

> *"There was a transition of the kind of object that we were looking at. Going from brighter objects to fainter objects so we could see further out. ... And when we went to the fainter objects, at lower signal-to-noise. It revealed problems in the pipeline that there were biases that you don't have at high signal-to-noise, but you do have at low signal-to-noise and it was just trashing everything."* (Astronomy data release manager)

Here this astronomer's explanation intentionally doesn't delve into the complex work undertaken by these collaborations, instead conveying that many of the nuances are fairly invisible to data users. Our subjects know changes are present between releases of their data but they tended to not have enough information to effectively evaluate how they impacted their work, at least until some part of their analysis infrastructure broke down. Documenting and surfacing the provenance behind these data changes is an aspect of this infrastructural work that is underdeveloped. Surfacing these changes through better tools and practices will become even more crucial to the longevity and utility of this essential scientific resource. Understanding types of change is a first step to doing so.

4.2 Two Broad Types of Data Change

While we find that data change is an issue for scientific work, investigating and designing for the issue in different scientific contexts requires first developing a

characterization of the notion. Our interviews foreground two general types of change for our subjects (1) change in the context and (2) a change in the data values themselves. These are not meant to be comprehensive or detailed, rather they're a first step at disambiguating what may be a highly variable concern among different disciplinary infrastructures.

Change in Context. Interest in changes to the context of datasets was consistently noted by subjects involved in managing data releases and archives, and sometimes by scientists using the files for their analyses. Such changes include: the organization of a data release's structure on the file system; the file naming scheme; the internal structure of files; and the metadata associated with the dataset, often encapsulated at least in part in file names and folder structures.

Data producers were particularly interested in unexpected changes to the context. These individuals are responsible for mirroring datasets across multiple computing systems for long-term storage and sharing. Verifying the consistency of the context, along with the data itself, is essential in this process. This work is difficult to easily do at scale with millions of files where thousands of changes may need to be assessed. It is also often invisible to most science end users of the datasets even though the results can impact their own work. Scientists were interested in context changes since they easily disrupt the operation of their software pipelines. They rely upon such pipelines to process and analyze their data. If changes to some structural aspect are not made visible they may encounter unexpected computational errors that waste valuable research time.

Change in Data Values. Scientists using data for analyses, as well as data producers managing releases, were concerned about changes to the data values stored within files. These end users indicate that they need to know not just that values have changed but importantly the amount of change. The magnitude of the changes influences how these users expect their analyses to be affected. This influences their decision process for further investigating the changes and potentially re-doing an analysis or resetting the starting conditions of a computational model they're building.

For example, earth scientists in our study use many streams of observational data collected at different sites, along with some satellite data, as input to computational models they're developing. Whether particular changes to data values matter significantly in this work varies, depending both on the amount of change and the specific type of data. One earth scientist explained such a case happening when the coordinates of a dataset were shifted by more than a meter. The project's cleaning process uncovered this mistake and disrupted the basis for gridding all of the data in their model. A data producer colleague adjusted the data being released, but their action was not readily visible to them or other end users. Our interviewee was informed about the change through their regular communication with this data producer and they had to go back and re-examine certain assumptions in the model, then re-execute it with the revised data. This scientist was effectively resetting their software instrument as a result of the change to the underlying data. Their work was influenced by the flow of the

project's releases, but they still expressed a belief that newer data releases were better due to the ongoing collection and cleaning work.

4.3 Trusting Collaborators and Processes to Make Datasets 'Better'

Collaborative projects release datasets with differing degrees of quality on varying timelines based on different factors, and with varying purposes. Our subject's work unfolds in concert with a changing web of relationships that interweave different artifacts, people, and practices. A science end user of a dataset may only be loosely connected to the infrastructural processes that created it. The details of the work can be a murky and invisible feature requiring trust in collaborators and their practices.

The work embedded within science infrastructures has effects which can shape decisions about which versions of data to use for analysis work. Exploring this, we wanted to know how our interviewees determine which version(s) of datasets to use for particular analyses given the fluctuating, evolving infrastructures and data releases. Our subjects reflexively stated that they use the latest data release available because it is the 'best' or 'better' than earlier versions. An exception was when the scientist knew that some data is no longer present in newer data releases, requiring they use an older version. Interviewees explained that they believe their data producer collaborators are always expanding their knowledge about the work and refining their practices. Astronomers continue to better understand their software pipelines, improving signal-to-noise ratios, creating cleaner and clearer images, and so on. Earth scientists remove bad data, fix sensors and instruments, and develop a longer record to base findings on. They trust that their colleagues are producing better products overall.

Our subject's trust in collaborators is closely connected to the organization of projects and their processes for producing data. Data production and cleaning is complex enough that no single person can fully follow every nuance of the work. For these astronomers the telescope and software pipelines build upon the long-term work of many researchers who develop deep knowledge of particular elements of the infrastructure's components. These earth scientists have little choice but to trust their colleagues who are directly connected to particular field sites and instruments, the individuals who can develop the strong tacit knowledge about this ongoing, remote work.

For example, an astronomy postdoc interviewed uses a numeric subset (rather than processed images) of the SDSS project's primary data release to develop statistical calculations of galaxy distribution. The postdoc has to trust colleagues who produce the overall release, as well as those creating the subset. This scientist won't know all of the subtle decisions that resulted in changes. Asked how they determine which version of data to use the postdoc replied "the latest" before explaining how they rely upon a chain of colleagues who are more hands on producing the numeric data subset. This individual is well aware of the complexity of the telescope and the software processing pipeline for removing bugs or systematics [13] and knows they can't reanalyze all of it themselves. Instead they

will call upon trusted collaborators when a bug arises and they lack the information needed to make an appropriate choice. A change may have little impact on their scientific analysis, or it may undermine the approach they're taking. At a glance the required knowledge is invisible and buried in the background.

> *So you need to have a close hand on the data to understand all of these potential systematics which could come in. So the people who create the datasets, which is not me. They know about this and I make sure that I'm using the latest datasets so that, I know that, that I have the, the best kind of dataset. ... And I don't do a re-analysis of the dataset. I, I trust the people who produce these datasets.* (Astronomy Postdoc)

This trust in data producer colleagues depends upon work embedded elsewhere in the project infrastructure. Change is expected in this iterative work, but the effects and particulars of changes are not readily visible to the end users of this infrastructure's key resource, even those who are members. Trust is essential to collaborative work [1,6,7,19], but the ability for an end user to verify and help identify issues if they have more information about changes to their data in between versions is important to further sharing of data widely outside the collaboration. Doing so may even take place outside of a project's established rhythms of data production and release.

4.4 Rhythms and Organizational Structures of Projects

Beyond trust, our subjects expect newer data releases to be better due to the organizational structure and community practices embedded in their research infrastructure shaping the rhythms of data production and release. The collaborative, multinational projects our subjects contribute to each gather, process, and release data with different timelines. This affects when versions of data are available to use in different forms of analysis work.

Astronomers in our study work on a yearly cycle that coincides with an annual weather pattern when observing is not possible. This data release team uses the time to wrap-up a year's data collection and get a release together. Some members of the project's different experiments (the way they organize different observing campaigns) have ongoing access to new data as it is collected since they are developing and refining software used to produce the final release. The eventual release, with re-processed data, is made available for members of the collaboration, then in time the public at large. For collaborators with some, or a lot of, visibility into this process they can influence the data by using them in preliminary analyses and reporting unexpected or incorrect effects. Their feedback can be folded back in to the data release team's work. End users not contributing to this process may eventually find changes in a new release and they may not have insight into the origin of these changes.

The earth scientists face a more fluid rhythm of data production, depending on the type of instruments and person(s) managing the flow of data from a field site to repository. Some PIs and groups may take many months or years

to gather and process data from one site before sharing with their larger collaboration. They develop and rely upon nuanced understandings of the physical context underlying their data and use this to process data to distribute within their collaboration. In other situations, the collaboration itself may directly manage instruments and the release of their data, applying standardized processing techniques and turning around new data within days, weeks, or months. Circumventing these rhythms, earth science subjects explained how at times they may have to go directly to a particular PI or instrument manager to get early access to data that has not fully been cleaned. This can be necessary when attempting to develop a baseline model of some system under study. In other scenarios, it is simply that a phenomena is so new that studying it requires rapid access to data that might not otherwise be available for years.

5 Discussion and Conclusion

We see that data change is an expected facet in scientist's work with inconsistent support for helping identify and address sources of change. Across our interviews and literature review we see that a focus of research infrastructures is producing and processing data for eventual use by scientists, typically project members but also in time a larger community too. Regardless of the timeline or rhythm to this work, important contextual information about changes to datasets is generated (whether explicitly or implicitly) and embedded within the enacted infrastructures. This information may not be readily conveyed in a visible manner to people beyond data production teams and a project's work practices may allow information underlying changes to fall by the wayside.

Surfacing concerns of data producers and science users trying to make sense of data change, one contribution of our study is to convey a general split between interest in types of changes in the contexts and in the data values, even as our subjects believe the latest data releases of a project are the 'best' available. Part of the challenge we as a community can address is determining what information it is possible to systematically produce to help end users of data products answer questions of relevance, trust, etc. Previous work that has explored trust and information seeking considerations [1,2,7,19,25] can be built upon to help scientists be able to more clearly understand and articulate why they find the latest data releases 'better' as their products evolve. At the same time, investigating non-computational provenance, as Thomer et al. [24] emphasize, along with provenance from particular computational workflows [22] is necessary.

Our findings offer a starting point for inquiries, even as there is more work to be done. Instead of treating data change as just a given facet of science we should continue to explore this realm as an opportunity for designing tools and practices to support scientists and help them grow and sustain their research infrastructures. We should design to support the capture and articulation of data changes that provide critical provenance, including quantitative information about its impact on downstream data analyses as well as qualitative insights. Our participant's projects (SDSS and Ameriflux in particular) currently do include

some information about changes between releases that is very general (e.g., new sites or observations added, major format changes, etc.). However, they do not provide much in-depth information that an end user scientist would need to assess whether they need to re-run analyses as they use the latest, 'best' data release. For example, changes in the filesystem may be encapsulated as part of the relevant contexts since many scientists in our study rely on folder structures or file names for at least some metadata in their work. In the moment, ephemeral information seeking leaves much of the labor less visible, if not invisible, to a variety of colleagues who take practices of the research infrastructure for granted. This iterative, ad hoc labor to identify and work with changes is another aspect to cleaning and processing data [14,16]. The resulting information produced about changes is really additional metadata about the scientific process itself that must be aligned to different contexts in spite of friction [6].

We see opportunities to design new tools and practices to help both end users of data and the collaborations producing releases since the process of working out and communicating changes between data releases is not well defined within the rhythms and organizational structures of our subject's projects. Currently, to effectively and appropriately use datasets scientists must undertake ad hoc, time consuming, iterative work to understand the product's structure and content, and differences from any past versions, among other concerns. Systematically designing tools and practices to surface data change should begin by supporting the work of users relying upon the data of research infrastructures. We can design tools to help calculate context changes so that data release teams—who are essential members of research infrastructures—can better communicate change information as a key element in their releases. Making visible their effort will furthermore help convey the care and craft that goes into change analyses.

In essence, designing to help construct information about data change means we are undertaking articulation work or metawork [8,23], ensuring that other researcher's (in this case science users) work can go well. Longer-term we should help communities develop common practices for explaining change in datasets, contributing to the sustainability of their research infrastructures. We can facilitate such efforts by building flexible software tools to integrate into different components of infrastructures and their shifting contexts. Infrastructure projects themselves should support and sustain these elements and produce change information as part of their data releases to aid their communities. There is a rich area of inquiry for design when investigating data change that has the potential to impact and shape a variety of research practices and facets of infrastructures at different scales. Shedding light on this work that is often invisible to end users is a first step in making such an impact.

Acknowledgements. The authors thank the members of the Deduce project, the study participants, and the anonymous reviewers of this work. This work is supported by the U.S. Department of Energy, Office of Science and Office of Advanced Scientific Computing Research (ASCR) under Contract No. DE-AC02-05CH11231.

References

1. Birnholtz, J.P., Bietz, M.J.: Data at work: supporting sharing in science and engineering. In: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, GROUP 2003, pp. 339–348. ACM, New York (2003). https://doi.org/10.1145/958160.958215
2. Borgman, C.L.: Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press, Cambridge (2015)
3. Dourish, P., Gómez Cruz, E.: Datafication and data fiction: narrating data and narrating with data. Big Data Soc. **5**(2) (2018). https://doi.org/10.1177/2053951718784083
4. Edwards, P.N.: A Vast Machine: Computer Models, Climate Data, and the Politics of Global. MIT Press, Cambridge (2010)
5. Edwards, P.N., Jackson, S.J., Bowker, G.C., Knobel, C.P.: Understanding infrastructure: dynamics, tensions, and design. Workshop report, University of Mighican (2007). http://hdl.handle.net/2027.42/49353
6. Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C., Borgman, C.L.: Science friction: data, metadata, and collaboration. Soc. Stud. Sci. **41**(5), 667–690 (2011). https://doi.org/10.1177/0306312711413314
7. Faniel, I., Jacobsen, T.: Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Comput. Support. Coop. Work (CSCW) **19**(3), 355–375 (2010). https://doi.org/10.1007/s10606-010-9117-8
8. Gerson, E.M.: Reach, Bracket, and the Limits of Rationalized Coordination: Some Challenges for CSCW Resources, Co-Evolution and Artifacts, Computer Supported Cooperative Work, pp. 193–220. Springer, London (2008). https://doi.org/10.1007/978-1-84628-901-9
9. Gitelman, L., Jackson, V.: Introduction. In: Gitelman, L. (ed.) "Raw Data" is an Oxymoron. Infrastructure Series, pp. 1–14. MIT Press, Cambridge (2013)
10. Jirotka, M., Lee, C.P., Olson, G.M.: Supporting scientific collaboration: methods, tools and concepts. Comput. Support. Coop. Work (CSCW) **22**(4–6), 667–715 (2013). https://doi.org/10.1007/s10606-012-9184-0
11. Karasti, H., Blomberg, J.: Studying infrastructuring ethnographically. Comput. Support. Coop. Work **27**(2), 233–265 (2018). https://doi.org/10.1007/s10606-017-9296-7
12. Kitchin, R.: The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences. Sage, London (2014)
13. Paine, D., Lee, C.P.: Who has plots? contextualizing scientific software, practice, and visualizations. In: Proceedings of the ACM on Human-Computer Interaction 1(CSCW) (2017). https://doi.org/10.1145/3134720
14. Paine, D., Sy, E., Piell, R., Lee, C.P.: Examining data processing work as part of the scientific data lifecycle: Comparing practices across four scientific research groups. In: iConference 2015 (2015). http://hdl.handle.net/2142/73644
15. Pipek, V., Karasti, H., Bowker, G.C.: A preface to 'infrastructuring and collaborative design'. Comput. Support. Coop. Work (CSCW) **26**(1), 1–5 (2017). https://doi.org/10.1007/s10606-017-9271-3
16. Plantin, J.C.: Data cleaners for pristine datasets: visibility and invisibility of data processors in social science. Sci. Technol. Hum. Values **44**(1), 52–73 (2019). https://doi.org/10.1177/0162243918781268
17. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. **23**(4), 3–13 (2000)

18. Rawson, K., Munoz, T.: Against cleaning. Curating Menus **6** (2016). http://curatingmenus.org/articles/against-cleaning/
19. Rolland, B., Lee, C.P.: Beyond trust and reliability: reusing data in collaborative cancer epidemiology research. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, CSCW 2013, pp. 435–444. ACM, New York (2013). https://doi.org/10.1145/2441776.2441826
20. Star, S.L., Ruhleder, K.: Steps toward an ecology of infrastructure: design and access for large information spaces. Inf. Syst. Res. **7**(1), 24 (1996)
21. Star, S.L., Strauss, A.: Layers of silence, arenas of voice: the ecology of visible and invisible work. Comput. Support. Coop. Work (CSCW) **8**, 9–30 (1999)
22. Stodden, V., et al.: Enhancing reproducibility for computational methods. Science **354**(6317), 1240–1241 (2016). https://doi.org/10.1126/science.aah6168
23. Strauss, A.: The articulation of project work: an organizational process. Sociol. Q. **29**(2), 163–178 (1988)
24. Thomer, A.K., Wickett, K.M., Baker, K.S., Fouke, B.W., Palmer, C.L.: Documenting provenance in noncomputational workflows: research process models based on geobiology fieldwork in yellowstone national park. J. Assoc. Inform. Sci. Technol. **69**(10), 1234–1245 (2018). https://doi.org/10.1002/asi.24039
25. Vertesi, J., Dourish, P.: The value of data: considering the context of production in data economies. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW 2011, pp. 533–542. ACM, New York (2011). https://doi.org/10.1145/1958824.1958906
26. Weiss, R.S.: Learning From Strangers: The Art and Method of Qualitative Interview Studies. The Free Press, New York (1995)