The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science

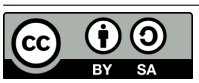
ANDREA K. THOMER, School of Information, University of Arizona, USA
DHARMA AKMON, Inter-university Consortium for Political and Social Research, USA
JEREMY J. YORK, School of Information, University of Michigan, USA
ALLISON R. B. TYLER, School of Information, University of Michigan, USA
FAYE POLASEK, School of Information, University of Michigan, USA
SARA LAFIA, Inter-university Consortium for Political and Social Research, USA
LIBBY HEMPHILL, School of Information, University of Michigan & Inter-university Consortium for Political and Social Research, USA
ELIZABETH YAKEL, School of Information, University of Michigan, USA

Data curation is the process of making a dataset fit-for-use and archivable. It is critical to data-intensive science because it makes complex data pipelines possible, studies reproducible, and data reusable. Yet the complexities of the hands-on, technical, and intellectual work of data curation is frequently overlooked or downplayed. Obscuring the work of data curation not only renders the labor and contributions of data curators invisible but also hides the impact that curators' work has on the later usability, reliability, and reproducibility of data. To better understand the work and impact of data curation, we conducted a close examination of data curation at a large social science data repository, the Inter-university Consortium for Political and Social Research (ICPSR). We asked: What does curatorial work entail at ICPSR, and what work is more or less visible to different stakeholders and in different contexts? And, how is that curatorial work coordinated across the organization? We triangulated accounts of data curation from interviews and records of curation in Jira tickets to develop a rich and detailed account of curatorial work. While we identified numerous curatorial actions performed by ICPSR curators, we also found that curators rely on a number of craft practices to perform their jobs. The reality of their work practices defies the rote sequence of events implied by many life cycle or workflow models. Further, we show that craft practices are needed to enact data curation best practices and standards. The craft that goes into data curation is often invisible to end users, but it is well recognized by ICPSR curators and their supervisors. Explicitly acknowledging and supporting data curators as craftspeople is important in creating sustainable and successful curatorial infrastructures.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing systems and tools*; • **Applied computing** → *Document preparation*; • **Information systems** → *Digital libraries and archives.*

Additional Key Words and Phrases: data curation, knowledge infrastructure, craft, coordination, workflows, social science data

Authors' addresses: Andrea K. Thomer, athomer@arizona.edu, School of Information, University of Arizona, Tucson, Arizona, USA; Dharma Akmon, Inter-university Consortium for Political and Social Research, Ann Arbor, Michigan, USA; Jeremy J. York, School of Information, University of Michigan, Ann Arbor, Michigan, USA; Allison R. B. Tyler, School of Information, University of Michigan, Ann Arbor, Michigan, USA; Faye Polasek, School of Information, University of Michigan, Ann Arbor, Michigan, USA; Sara Lafia, Inter-university Consortium for Political and Social Research, Ann Arbor, Michigan, USA; Libby Hemphill, School of Information, University of Michigan & Inter-university Consortium for Political and Social Research, Ann Arbor, Michigan, USA; Elizabeth Yakel, School of Information, University of Michigan, Ann Arbor, Michigan, USA.



2573-0142/2022/11-ART414
https://doi.org/10.1145/3555139

ACM Reference Format:
Andrea K. Thomer, Dharma Akmon, Jeremy J. York, Allison R. B. Tyler, Faye Polasek, Sara Lafia, Libby Hemphill, and Elizabeth Yakel. 2022. The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 414 (November 2022), 29 pages. https://doi.org/10.1145/3555139

1 INTRODUCTION

Data curation—the work of making data fit-for-use, archive-ready, and accessible over the long-term—is critical to data-intensive science [10, 24, 35, 61, 91]. In data science contexts, this work is often referred to as munging, wrangling, or processing, with a particular focus on getting data into a usable format. This process of making data fit-for-use can take up to 80% of a data scientist's daily work [86]. In institutional contexts—for instance, in large scientific data archives or institutional repositories—data curation is likely to involve the application of data and metadata standards in addition to data munging, with a particular focus in making data shareable and easy to reuse. In both contexts, the ways in which data are transformed and manipulated prior to analysis have significant impacts on the quality and reliability of a study [9, 19, 56]. Additionally, making data ready to archive or share is increasingly required by both funding agencies and journals.

Despite its importance, data curation—and data curators—are often overlooked in accounts of data science. Job ads for data scientists frequently call for data "unicorns," "ninjas," and "rock stars" to wrangle messy datasets using mythic abilities (not through skill or craft) or "janitors," as if data processing were a rote sanitizing process that requires little specialized expertise or training [19]. Data science clients similarly think of data work as "magic" [46]—which, while seemingly complimentary, elides the skill, effort, and careful decisions that go into data curation and that ultimately impact the trustworthiness and reproducibility of a study.

The work of data curation can also be obscured, somewhat ironically, through attempts to render it visible as part of a regularized workflow. Workflows and curatorial best practices aim to break curation into a discrete set of steps, or show it as one "phase" of work in a project (e.g., [36, 53]). The goal of workflow representations is to make curation more reproducible and routine, but it comes at the cost of obscuring the skill needed to do these tasks well and furthering the idea that curatorial work is a rote task that just any human can be plugged into. In this way, curatorial work can be viewed as akin to factory labor, in which data workers are interchangeable parts of an assembly line [65].

Obscuring data curation also renders the labor and contributions of data curators invisible [64]. Like other forms of service work, well-executed curation is hidden [77]. Additionally, obfuscating curatorial work makes it challenging to understand the impact of specific curatorial actions, and therefore to efficiently prioritize, plan, or fund data curation [33]. Without understanding the impact of data curation, the developers of curatorial tools cannot assess or prioritize which features and functionalities will best increase curatorial efficacy or later data reuse.

To make the work of data curation more visible, we conducted a close examination of data curation at a large social science data archive, the Inter-university Consortium for Social and Political Research (ICPSR). ICPSR recently adapted external standards and best professional practices to create robust internal guidelines for curation, and the scale, centrality, and collaborative aspect of curatorial work at ICPSR make it an excellent site for a case study of data curation. ICPSR is the largest social science data archive in the world, and it contains datasets from over 16,000 studies. ICPSR's professional in-house curation activities distinguish it from other data repositories such as the UCI Machine Learning Repository[1] or Dataverse[2] where data providers are expected to prepare

[1]https://archive.ics.uci.edu/ml/index.php
[2]https://dataverse.harvard.edu/

data for sharing and preservation themselves. This research is part of a larger project focused on understanding the impact of curatorial work and aimed at developing metrics that better measure and account for the benefits of that work [33, 47]. We aim to make curatorial work more visible and thereby easier to account for in budgets and in academic promotion cases. Our methods include interviews with ICPSR stakeholders, as well as computational analysis of curation logs. We address the following research questions:

(1) What does curatorial work entail at ICPSR, and what work is more or less visible to different stakeholders and in different contexts?
(2) How is that curatorial work coordinated across the organization?

We drafted these questions with the goal of understanding both the visible and invisible work that goes into data curation; prior studies have shown that much of this work escapes view [64], but few have sought to specify what, exactly, is invisible.

We found that although there are several standard curatorial activities performed at ICPSR, and well-defined standards for different "levels" of curation, considerable craft and coordination are needed to enact these best practices. In other words, craft is needed to "work" a workflow, and nontechnical work is necessary to facilitate technical work. Surfacing the role of craft and coordination has important implications for curatorial projects' and teams' planning. This defies the rote sequencing of events implied by many life cycle or workflow models. We provide a detailed account of curatorial work at ICPSR and explain how workflow-centric accounts of data curation can obscure both the individual skilled "artistry" and coordination necessary in this work. In doing so, we bridge research in Computer-Supported Cooperative Work (CSCW) and the library and archival sciences on data curation.

We additionally reflect on the visibility of data curation, both within ICPSR and to data users. As Plantin [64] has previously described, much of data curators' work is intentionally kept invisible to the final data consumers—yet curators experience their jobs as being hypervisible to their supervisors, via the extensive documentation they create. We discuss how different kinds of invisible work are at play in data curation at ICPSR and explain how CSCW and data science can benefit from better understanding the judgements and skill that goes into effective data curation.

2 PRIOR WORK

2.1 What is data curation?

Data curation has multiple definitions and multiple parallel lineages [58]. In this subsection, we briefly review definitions of, and research on, data curation in the information sciences (IS) as well as in CSCW. We take a broad view of what constitutes data curation partly because the data archive at the center of our study, ICPSR, conducts a broad range of curatorial work—including direct manipulation (akin to "wrangling" or "munging") of data, the creation of metadata, and the long-term management and archiving of data (a detailed description of data curation activities at ICPSR is provided in Section 3.1). However, we additionally take this broad definition of data curation to align with the majority of prior research in IS, and thereby show overlaps between research in IS and CSCW.

IS scholars and practioners define data curation as the "the active and on-going management of data through its life cycle of interest and usefulness to scholarship, science, and education" [14] (see also [47, 61, 91]).[3] Research in data curation has included substantial scholarship on data practices

[3] "Digital curation" is often used as a broader, more general term for management of any collection of digital objects [91], whereas "data curation" refers to the long-term care and management of data specifically. Other terms used for this work include research data management, data stewardship, and digital preservation. We use the term data curation in this paper but draw from research on digital curation and preservation where relevant.

(the ways in which people work with, share, and reuse data; examples include [15, 76, 81, 92, 97]) and on understanding the full "life cycle" of data use (e.g., [4, 37, 79]). This latter genre of research has resulted in numerous data or data curation life cycle models, such as the Digital Curation Centre's Curation Lifecycle Model [36] and the Big Data Lifecycle Model [66], which identify different phases of data work from the point of data collection through its publication or archival. Some life cycle models focus primarily on curatorial workflows (e.g., Higgins [36]), whereas others describe the broader cycle of data work in science and scholarship (e.g., Feger et al. [26]). Models of curation might also take the form of terminological frameworks, such as the "Data Practices Vocabulary" by Chao et al. [12], which outlines a taxonomy of terms describing curatorial work.

Libraries, data archives, and other repositories conduct a range of curatorial work throughout the entirety of the data life cycle. In some cases, curators can begin actively curating data early in the data life cycle alongside the data producers; this is particularly common in large, domain-specific archives like ICPSR or those run by scientific agencies like NASA or the United States Geological Survey. However, in smaller institutions, curators may primarily engage with data after deposit. Thus, depending on the context, data curators work with data at all levels—from the variable level to the dataset level through activities such as applying metadata at multiple levels, checking for consistency, and ensuring that the documentation accurately represents the data [66].

Although much of the CSCW literature does not discuss data curation *per se*, it does discuss hands-on work with data with the goal of improving its fitness for use by a single user [26, 42, 78]. For example, Feger et al. describe on the activities "essential for generating reproducible artefacts" via research data management, whereas Kandel et al. are concerned with "wrangling" data to enable meaningful analysis for the research at hand. Others, such as [53] identify curation as one type of human "intervention" that results in data for analysis and includes data cleaning, metadata conversion, and data alignment. In proposing "datasheets for datasets," Gebru and colleagues [Gebru et al. [28]] argue for a key aspect of curating data for reuse: that data used in machine learning should carry documentation that describes its collection processes and transformations. For data science workers, data curation is a collaborative activity centered on information exchange and data and code transparency [94]. In their ethnographic study of the Long Term Ecological Research Network, Karasti et al. [44] describe "information managers"' data stewardship strategies and strengths, including their ability to turn localized, heterogeneous data into a networked resource. In this way, data curation and the development of information infrastructures are long-term sociotechnical endeavors. Finally, the growing literature on data work (e.g. [63, 70]) and data practices (e.g. [15, 76, 81, 92]) are both concerned with the ways in which people gather, manipulate, share, and otherwise interact with data.

2.2 What renders data curation invisible?

What may not be clear from all these definitions of curation work is that when curation is done well, it is invisible to the data's users. Successful data curation enables reusers to readily access and use datasets and does not highlight the data transformations or metadata generation that make the data ready for use. Curation work can be done by the data producers themselves as part of their research data management process. This curation work includes cleaning, organizing, and storing their data for their own localized research needs [85]. Often, however, curation tasks—the data manipulation, cleaning, documentation, preservation, and other work discussed below—are carried out by data curators or processors within the repository or archive that has selected the data for inclusion in its holdings [41].

The role of the data archive is two-fold: to enable the researcher to share their data with the scientific community and to ensure the data's long-term accessibility and preservation [32]. The data producer deposits their data with the archive, and then at some future moment, the data appear

in a standardized form, ready for use, with inconsistencies smoothed over and issues addressed, presented to the data user without the explicit traces of the curation work, which are only visible to those within the data archive [45, 64]. By producing data according to professional standards, for instance, curators purposefully render themselves and their work invisible [64].

The invisibility of the work makes it hard to classify. Prior scholars have explained myriad ways that work can be invisible. For instance, Nardi and Engeström [55] identify four types of invisibility at work: 1) work done in invisible places, such as the highly skilled behind-the-scenes work of reference librarians; 2) work defined as routine or manual that actually requires considerable problem solving and knowledge, such as the work of telephone operators; 3) work done by invisible people, such as domestic workers; and 4) informal work processes that are not part of anybody's job description but which are crucial for the collective functioning of the workplace, such as regular but open-ended meetings without a specific agenda, informal conversations, gossip, humor, and storytelling.

Similarly, Star and Strauss [74] propose three forms of invisible work: where "the act of working or the product of work is visible to both employer and employee, but the employee is invisible"; where the "workers themselves are quite visible, yet the work they perform is invisible or relegated to a background of expectation"; and when "both work and people may come to be defined as invisible," according to particular indicators. Curators possess different types of invisibility. For example, D'Ignazio and Klein [19] recognize the highly skilled behind-the-scenes work prevalent in what they term "data cleaning" and at the same time recognize that others discount the intellectual work required. Kross and Guo [46] report on the black-boxing of curation that leads clients to deem the results "magic."

In social computing, curation work is sometimes invisible because it occurs during data collection or generation. Machine learning, computer vision, and social media studies often use "found" data [33, 39, 62] and render curatorial decisions such as "what data should be available," "in which format(s) should data be provided," or "how should this data be sampled" invisible. For instance, datasets scraped from the web (such as Flickr photos [71, 95] or Wikipedia talk pages [89, 90]) suffer from biases in representation [39]. The kinds of curation activities that occur in archives could address those biases by adjusting samples, weights, or documentation. Annotation processes in which humans add labels to data that can then be used in machine learning tasks (e.g., facial recognition, hate speech detection) are another type of data generation step that is often minimized in reports about the research that depend on them. For instance, Scheuerman et al. [71] explain that reference datasets used in computer vision tasks in papers are not well-described, and the details of the annotation process and potential biases introduced are missing. They argue that the value of "efficiency" is responsible for this pattern and that explicitly working toward other values such as "care" could improve data curation practices in computer vision.

2.3 Craft in data work

Throughout the CSCW literature, there is discussion of the craft needed in technical work; recent papers have begun to apply this framework more specifically to work with data. Definitions of craft are varied; Merriam-Webster defines craft as "skill in planning, making, or executing" [50]. Richard Sennett takes a likewise broad approach that considers a variety of skilled work, from carpentry to programming, as craft; craft is less making a thing than making a commitment to do a job well [73]. Dormer notes that craft has changed meaning, evolving from a description of acumen and a way of doing, rather than making, to one that focuses on making [20]. The CSCW literature has largely shied away from firm definitions of craft, preferring instead to describe attributes of craft. Barley and Orr [5]'s well-known volume on the topic argues that "technical work sits at the intersection of craft and science, combining attributes of each that are normally thought to

be incompatible." The attributes of craft include the use of complex technologies, a reliance on contextual knowledge and skill, the development of abstract conceptual representations to guide work, and a grounding in a community of practice [6]. Rosner et al. [68] argue that appreciation of craft work in computer science, though, is marred by "gendered narratives" about the value of such labor, which, "both haunt and inform HCI's ideas of technological belonging, participation, and differentiation." Cheatle and Jackson [13] focus on the increasingly digital aspects of craft and roles of creativity and collaboration in craft work. In the context of data work, scholars have focused on how craft practices are used to process and interpret data. Mentis et al. [49] examine surgeons collaborating remotely over image data to craft a shared interpretation. More recently, Muller et al. [53] view the data science pipeline process as one of that crafts the data, a process in which workers combine technical skill, expertise working with abstraction, and representations and decision-making to accommodate unexpected issues and application of more routine techniques and automated scripts.

Within IS, discussion of craft has largely focused on its role as part of librarianship and archival practice. Archivist Trevor Owens brings these conversations forward to a digital context in his discussion of digital preservation (an aspect of data curation) as a craft, which is "best understood as part of an ongoing professional dialog on related but competing notions of preservation that goes back to the very beginnings of our civilizations" [58]. He further writes:

> Digital preservation must be a craft and not a science because its praxis is; 1) grounded in an ongoing and unresolved dialog with the preservation professions and 2) it must be responsive to the inherent messiness and historically contingent nature of the logics of computing.

Given the broad definitions of craft and the assertion of craft work in both data science and digital curation, we have extended these lines of thought to include curators as craftspeople. In particular, they exhibit the ability to engage with abstractions of work, creativity, and collaborative processes detailed in the research on craft.

When this craftful work is done collaboratively, considerable articulation work and coordination are needed to do it successfully. Articulation work "consists of all the tasks needed to coordinate a particular task, including scheduling subtasks, recovering from errors, and assembling resources" [29], whereas coordination is the "process expertise" entailed in said scheduling and assembly [6]. Articulation and coordination work have been shown to be critical in data curation for multiple reasons. They are needed in maintaining a knowledge infrastructure's stability [43], facilitating the selection and enactment of data curation protocols [16], refactoring data structures and vocabularies [82], supporting infrastructure design [3], enabling navigation of information during the process of scientific discovery [59, 60], and they are a core component of the "data labours" of building and sustaining data collections [54]. Erickson and Jarrahi [22] describe the articulation work needed by knowledge workers, such as data curators, to configure infrastructural solutions to overcome technical and contextual constraints in tools and workplaces. A recurrent theme in these papers is the lack of tools to support this coordination and articulation work; curators must coordinate their work often in spite of these tools rather than through them.

3 METHODS

In this paper, we report on a mixed methods study to examine different aspects of the data curation process. We leverage two bodies of data: 1) semistructured interviews with stakeholders across ICPSR and 2) records of curation work in Jira tickets, a subset of the internal ICPSR documentation that records data curators' work.

3.1 Research Site

ICPSR, founded in 1962, is one of the oldest and largest curated social science data archives in the world. It not only collects, curates, and disseminates data in a broad range of disciplines—including political science, sociology, demography, education, criminology, public health, among others—but is also a leader in repository infrastructure, data curation standard setting, and innovation in data curation. ICPSR's archives include over 16,000 studies that contain nearly 6 million variables. ICPSR's collections are organized into separate archives representing different subject areas and are often sponsored by federal agencies and foundations. We selected ICPSR for three reasons: 1) their curation processes are well articulated and documented, 2) the volume of data curation is large enough for patterns to emerge, and 3) we were given access to both documentation and staff to conduct an in-depth study of the curation process.

ICPSR's organizational structure also makes it possible to study data curation in depth. Several years ago, ICPSR centralized curation into one unit. Curators previously worked for individual archives within ICPSR, reporting to a project manager, who, in turn, reported to an archive director. Now, curation staff, project management staff, and archive directors sit within their own distinct organizational units (e.g., the Curation unit, the Project Management unit). Part of this reorganization also involved a redesign of curatorial standards. As of 2018, datasets are assigned to one of three standard "levels" of curation that articulate specific curatorial actions that vary according to the amount, intensiveness, and complexity of effort required as well as the end product delivered. These levels provide a standard for curation actions and expected outputs, which are assigned based on the format, size, and level of preparation performed by the data creator prior to deposit [38]. Higher levels of curation are intended to improve the usability of data products. All data deposited with ICPSR receive a base level of curation ("Level 1 Curation"), meaning that curators remediate personally identifiable (disclosive) information and create a metadata record, a Digital Object Identifier (DOI), statistical files, a web page, and a codebook explaining the variables in the data collection. "Level 2 Curation" includes all Level 1 actions plus additional data transformations, completeness checks, and preparation of the data for online analysis. "Level 3 Curation" is intensive and includes custom documentation, attached survey question text to variables, and indexing variables for search. Nontablular data, such as qualitative or spatial data, typically require Level 3 curation. For example, the "TransPop, United States, 2016-2018" study shown in Figure 1 is curated at Level 3, meaning that additional curatorial tasks have been assigned and more time has been budgeted for intensive curation, including extensive disclosure review and remediation and creation of searchable question text.

3.1.1 Interviewees and semistructured interviews. The internal stakeholders in ICPSR curation extend beyond the curation unit itself. In order to better understand the impact of curation within the data repository, we conducted in-depth, semistructured interviews with 37 ICPSR stakeholders comprising 6 staff groups: archive directors, project managers, curation supervisors, curators, user support, and bibliographers. Each archive is led by a director who spearheads collection development efforts, secures funding, interacts with archive sponsors, and attends disciplinary conferences and meetings to expand the reach of the archive. When ICPSR ingests data, a project manager shepherds the data through curation and dissemination, serving as a conduit between the curators and the data producers. The project manager works with the archive director and the curation supervisor to determine which curation activities should be applied to the data and how to prioritize the data relative to other studies in the queue. User support personnel are the bridge between data reusers and either project managers or curators as questions about the data arise. Bibliographers track use of all of ICPSR's curated datasets and maintain an extensive bibliography of that use.

Curation work is accomplished primarily by a dedicated team of curators (n = 32) and their curation supervisors (n = 5). We note that data curators are typically entry-level employees; this is not always the case at data archives. ICPSR requires curators to have experience with statistical software (e.g., SPSS, Stata), data preparation, and social science research methods. ICPSR actively curates data to ensure that they comply with the FAIR principles (i.e., are findable, accessible, interoperable, and reusable) [87]. Generally, curators review data for sensitivity and reidentification risk, generate metadata [84], identify missing values, index variables for future search and discovery, link question text to variables, apply subject terms to the study, and generate the data files in multiple formats (e.g., SPSS, Stata, plain text). Curators also pass along citations to publications that use the data to the bibliographers for inclusion in the ICPSR Bibliography of Data-Related Literature. These tasks are completed by individual curators and reviewed by their supervisor or a senior curator before disseminating the data for reuse. On an ongoing basis, the bibliographers also search for additional citations to studies archived at ICPSR.

The interviewees were selected using purposive sampling [51]; we requested interviews with all personnel working in the specific roles identified. The only criteria used to filter out potential respondents was for the curators themselves: due to the time required to become familiar with curatorial work, we limited our interview requests to those curators who had been working for at least one year so that they had built up some expertise in curation work and could better reflect on the processes. The interviews focused on understanding how the different stakeholder groups measured the value and impact of curation work (see Appendix A for our full interview protocol).

We conducted 37 semistructured interviews with archive staff that enabled us to probe further into the responses and ask questions that were specific to each role [18, 34, 69]. Interviews were conducted in 2019 and 2020; 12 were face-to-face interviews conducted before the COVID-19 pandemic led our institution, the University of Michigan, to transition to remote work, and the remaining 25 were conducted remotely via Zoom and Google Meet. Table 1 details our interview participants. The interviews were recorded and then transcribed by the REV transcription service and verified. We anonymized our interview transcripts by assigning identifiers to all participants. Our study was reviewed by our university's Institutional Review Board and found to be exempt from ongoing oversight.

We began analysis using a deductive approach, using a "start list" of codes derived from our research questions, interview questions, and our knowledge of prior literature. In this first round of codes, we paid particular attention to identifying curatorial actions at ICPSR. Codes were were iteratively expanded and refined through subsequent rounds of inductive coding [48, 51]. Analysis of the interviews was completed using the qualitative data analysis program NVivo. Because multiple team members were conducting the coding, we establish interrater reliability (IRR) to ensure coherence. As we began establishing IRR between two members of the interview team, we realized that the interviews between the different stakeholders were divergent enough that IRR would need to be established within each stakeholder group. With the exception of the single User Support interview (59.2%) and the three Bibliography Team transcripts (69.5%), IRR was repeated within each stakeholder group until at least 70% was achieved using Scott's pi [72]. One member of the interview team coded transcripts across all stakeholder groups, and two different members established IRR with her on specific sets of transcripts.

After each round of coding was completed, team members reviewed coded data, then met as a group to discuss emergent themes. After our first round, we identified the role of craft and coordination as being key to data curation work and decided to conduct a secondary round of axial coding to deepen our analyses. Again, IRR was established between team members until at least 70% was achieved using Scott's pi. The authors reviewed coded data and again met to discuss the codes as a group. Finally, after reviewer feedback, we conducted a third round of coding, this time

Table 1. Number of interviews by stakeholder category

Stakeholder Category	Number of Interviews	Interview Codes
Archive Director	7	AD002-AD006, AD008, AD011
Project Manager	9	PM025-PM033
Curation Supervisor	7	CS001, CS007, CS009, CS010, CS012-CS014
Curator	10	CU015-CU024
Bibliography Team	3	BT034-BT036
User Support	1	US037

diving deeper only into codes related to craft in data curation to deepen our analysis. We met again as a group to discuss emergent themes.

3.1.2 Triangulating with Jira tickets. We triangulated findings from interviews with documentation created through the curation process, again looking for descriptions of curatorial actions. Curation work at ICPSR is coordinated and documented across three main sets of documents: processing plans, Jira tickets, and processing history files. Jira tickets are the richest and most specific record of data curators' work. Jira is type of project management software that organizes work through the creation of "tickets" that describe the work that needs to be done and that users can update with progress over time. When data are deposited through the ICPSR deposit system, staff review the data for fit and priority, and a data project manager or assistant generates a Jira ticket (see Figure 1; we removed identifying information from the fields on the right but left their titles to show what information tickets contain). These tickets provide a study title, the priority of the study, the funder or sponsoring archive, a description of the work curation will need to do, and the level of curation (and any additional tasks) required. Before curation begins work on a Jira ticket, Metadata Unit staff review the ticket and study the metadata. After data project staff and metadata staff approve the ticket, it is sent to curation for assignment. While curation works on the study, they provide details about their work and progress in the "worklog" section of the ticket (see Figure 2). The worklogs offer insights into aggregate time spent on different kinds of curatorial actions at ICPSR.

To classify the parts of worklog descriptions (e.g., "Began curation" and "Metadata and proc plan" in the example in Figure 2), we developed a set of eight high-level curatorial actions that describe curation work: initial review and planning, data transformation, metadata, documentation, quality checks, communication, noncuration, and other activities (see Figure 3). These categories mirrored the codes used in our qualitative analysis.

We manually coded a randomly selected proportional sample of Jira ticket worklog entries stratified by curation level. These were coded in *brat* software [75] to create labeled training data to facilitate the automatic classification of the Jira ticket worklogs (discussed more fully in Lafia et al. [47]). We trained a computational model with 0.75 accuracy to assign each worklog entry one of the eight categories of curatorial actions (summarized in Figure 3). For example, a worklog entry "Discussed curation standards with supervisor (2 hours)" is classified as an instance of "Communication," whereas an entry like "Recording dataset limitations in processing notes (10 hours)" is classified as "Documentation." We then aggregated each class of action to analyze the relative amount of time spent on each.

There are several limitations to this study. First, it documents curation work at one repository. Second, ICPSR is a mature repository with well-articulated policies and procedures. Finally, we did not directly observe the curatorial process but relied on direct reports from curation staff and other stakeholders and indirect observation through the Jira tickets.

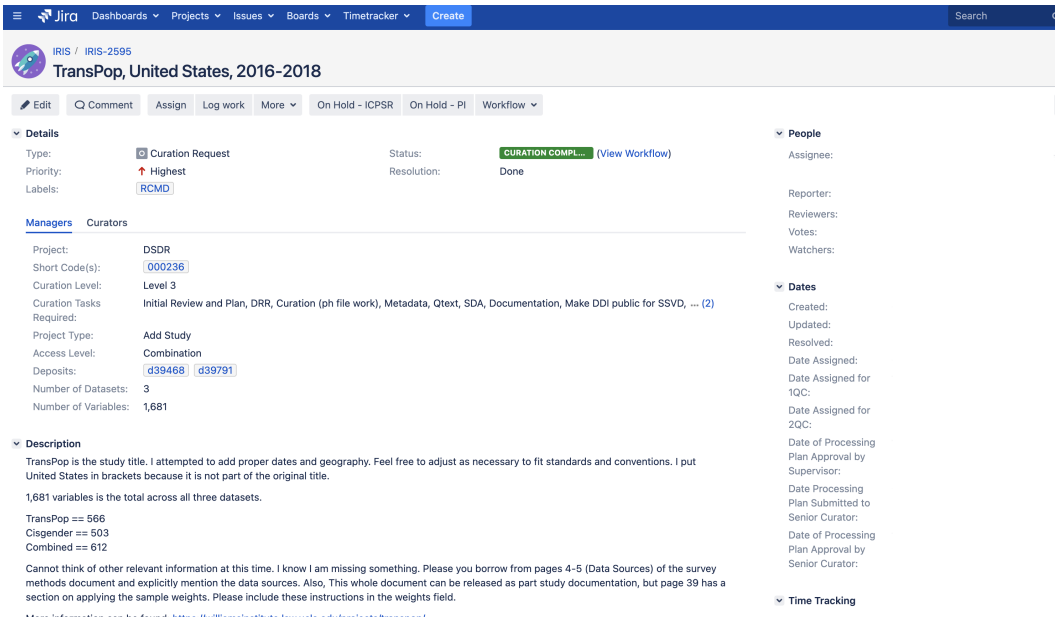


Fig. 1. Jira ticket for a single study

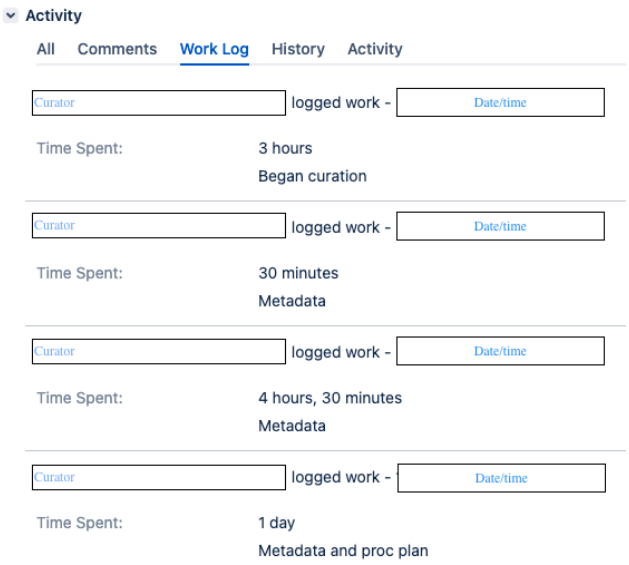


Fig. 2. Worklog excerpt from a Jira ticket

4 FINDINGS: CURATORIAL WORK AT ICPSR

In the interviews and Jira tickets, we found a consistent, overlapping vocabulary of actions describing typical curation work. We also found insights into the ordering and time spent on curation tasks (see Figure 3 for examples of each type of action). Table 2 summarizes the amount of time

curators logged for each type of curatorial action. While there were some typical sequences in which actions are performed, curators describe considerable variability in their own day-to-day work, due to their specific preferences, expertise, and craft knowledge.

In the subsections that follow, we first describe the core high-level curatorial actions that are undertaken at ICPSR. These expand prior accounts of the work of data curation in CSCW and data science, where it's described as one small part of a process. Notably, ICPSR data curators do more than simply catalog datasets as they come in; they work directly with datasets to "wrangle" them into usable, shareable, and archive-ready shape. They also create extensive documentation and metadata to help future users of this data. After describing the work of data curation at ICSPR, we describe how curators rely on their craft knowledge to navigate the workflow dictated by these actions. In doing so, *we show how best practices and craft practices are deeply intertwined.*

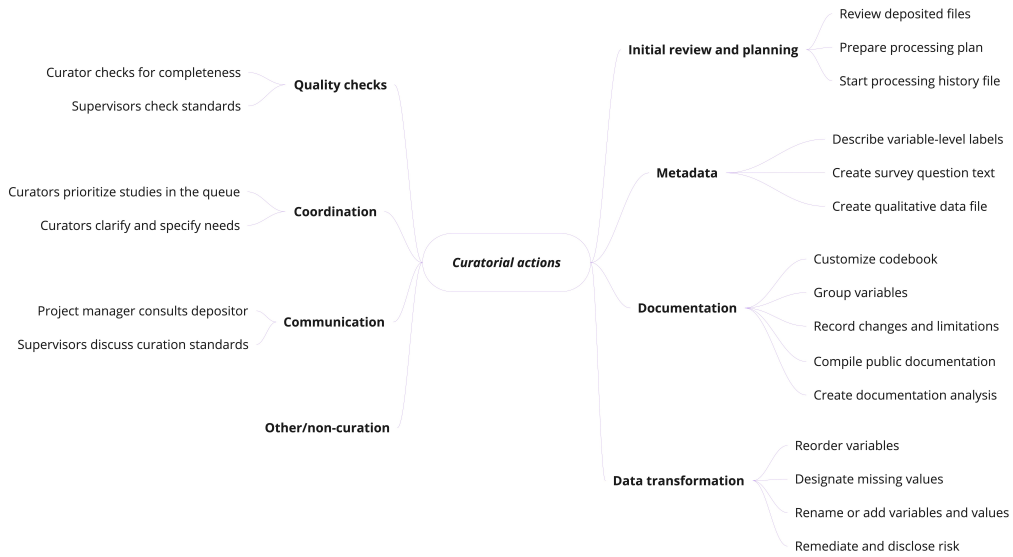


Fig. 3. High-level curatorial actions that occur throughout the curation process

Table 2. Time spent on curation actions (Feb. 2017 - Dec. 2019)

Action	Total hours logged	Percentage
Communication	3,249	7%
Data transformation	12,363	26%
Documentation	3,094	6%
Initial review and planning	5,778	12%
Metadata for study	2,669	6%
Noncuration	6,641	14%
Other	1,157	2%
Quality checks	13,075	27%

4.1 High-level curatorial actions

Initial review and planning. Data curators at ICPSR typically begin their curation of a deposit by reviewing deposited files and metadata and *developing a processing plan*—an outline of planned curation tasks, depending on a dataset's designated curation level. More detail about curation levels at ICPSR can be found in Section 3.1.2. These plans are developed by curators and reviewed by curation supervisors, who answer questions, troubleshoot, and generally advise along the way. In recent years, more initial review and planning actions have been recorded for higher levels of curation (Levels 2 and 3 Curation), suggesting that relatively more attention may be dedicated to developing curation plans at higher levels of curation. Initial review and planning accounted for 12% of curation time over our study period.

Early curation work also includes *disclosure risk review (DRR),* in which curators evaluate the risk that publishing a dataset might pose to research participants and identify appropriate mitigation steps. Curators described DRR as a critical way in which they add value to a dataset—both because of the anonymization it provides and for the thorough oversight the DRR represents.

Data structuring and transformation. This category of curatorial work includes the most "technical piece" of curation [CU015] akin to the "design" and data shaping described by Feinberg [27]: the direct work with the dataset itself to make it easier to use, share and archive. Data transformation tasks include designating missing values (e.g., assigning metadata to values like "no response" and "not asked"); adding question text (inserting the survey questions verbatim); transforming curated SPSS data files into other statistical packages (e.g., R, SAS, etc); and creating documentation (PDF codebooks and XML metadata files). Datasets are sometimes split into multiple, more usable parts (for instance, smaller file sizes or commonly used file formats) or are merged into single files from multiple sources. This data structuring entails more than just mechanical reformatting; as one curator describes, "a lot of times, especially in the larger datasets, there's a lot of pieces to put together that I think when we make those connections it makes it easier for users to use the data." [CU023] Considerable expertise and judgement are needed to structure and transform data well. Data transformation was comparatively time consuming, taking up 26% of curation time over our study period.

Metadata creation and improvement. Curatorial work includes the creation of records that will be queried by users within ICPSR's online repository. Metadata development is seen as a distinct task: where the focus of data structuring is to make the dataset usable in and of itself and the focus of metadata creation is to support search and retrieval of datasets. Curators saw metadata as particularly important because it's "the first line" of access [CU017], the first thing users see. Metadata improvements include drafting or revising a dataset's description; copying and refining metadata from the initial data provider, such as data collection dates; creating question text (i.e., writing out the full list of questions in the survey instrument that generated the data); and defining variable-level labels (i.e., creating a data dictionary that spells out what each data variable represents). Metadata work accounted for 6% of curation time. This work is both qualitative and technical; curators must have skill manipulating metadata standards to create these records, and they need to have the experience to understand what context to include in metadata records that is necessary and helpful for data reusers.

Documentation creation and improvement. In addition to creating metadata records, curators also develop other forms of documentation about the datasets. This includes creation of processing history files; codebooks, which include information for each variable in a dataset; documentation of major changes made to the data; and compilation of any additional documentation archived by the data producer. Documentation accounted for 6% of curation time logged. Curation activities in topical archives generated more recorded documentation that those in the general archive.

Quality checks. These include checking data and metadata files for completeness, confirming that the work done to a dataset aligns with the Jira request, comparing the work done to the processing plan, and confirming adherence to ICPSR's guidelines and protocols for curation. The vast majority of studies include quality checks, which was the major category of curation action we detected in our analysis of Jira ticket worklogs and accounted for over 27% of curation time logged. These quality checks are performed by a second curator to provide an extra level of review.

Beyond designated quality checks, stakeholders discussed the value that curation provided in ensuring that the data was of high quality overall. Project-related communication is one mechanism for ensuring high-quality curation and accounted for about 8% of curation time logged in Jira tickets. This includes catching issues and addressing complicated data challenges that data producers did not, providing high-quality documentation about data and the data curation process, setting and meeting goals for data release that match depositor expectations and deadlines, and being a source of consistent, vetted data. A curator described the last item in this way:

> Our work, I think, is pretty impactful and benefits the community because the work we put in [will] rule out all the troubleshooting. We look at the data, compare it to the documentation, and then do these things to make sure everything's consistent. If there's any problems, we either resolve with the PI or we have our own solution for it, so once you get your hands on the data, there's nothing really in question for the most part. [CU018]

4.2 Using craft to work the curatorial workflow

The previous section outlined the tasks involved in the technical work of data curation. Actually accomplishing that work, however, requires craft. Each curator aims to craft a clean, normalized dataset with standardized metadata and easy-to-read documentation—but the path to this final outcome is different for every dataset. Specifically, curators organize their work by *first developing a gestalt, abstract mental representation of the data to envision what the final released dataset will entail*; they then *use their judgement and expertise to interpret standards, creatively come up with solutions*, and thereby achieve a standard outcome in unstandardized ways. Paraphrasing one curatorial supervisor, "I'm the curator: I do anything needed to make the data archivable" [CU015]. We describe these two aspects of curatorial craft practices in the following subsections.

4.2.1 From abstract representations to fit-for-use. Curators approach a new dataset by getting the gestalt of the dataset: understanding the whole of the dataset as beyond the sum of its parts, as well as how these fit together in order to assess the feasibility of the processing plan and to envision the archivable and disseminated dataset. Several curators described getting the gestalt:

> I don't ... I do find the plan useful and it has to be done and there are things that it walks you through that can help you find missing things or problematic things. But I also rely heavily on just running the frequency output on the data and just scrolling through it. ... But I have found, many times I have found things there that I wouldn't have seen otherwise. So I find that very important. And then I just ... yeah, between the plan and this check on my own, I find out what I need to do with the data and the documentation ... I'm not always very linear in how I work with this. I tend to do a lot of poking around... (CU017)

> So myself personally, I'll try to work with the data first, make some data manipulations or changes after I've gone through and read the documentation that's been provided, and get a good sense of what's going on with the study. But I won't... So I'll read through everything and I won't fill out the metadata at that point because I still like to

> be able to go through and actually work with the data before I fill out the metadata that explains the collection in some more detail. [CU023]

Much of the process of getting a sense of the data is done with the user in mind. Curators think about how the dataset would need to be structured or documented in order to be fit-for-use for a range of users. Multiple curators describe customizing their work, or making decisions with the goal of supporting users' access, essentially envisioning themselves as the "first user" of a dataset and "trying to figure out everything that a potential user would want to know and to make sure the archive version that we release is as complete as possible" [CS009]. The curators' goals are to answer any questions future users might have about the data: "We obviously can't anticipate everything, but we try to say, okay, if I was just picking this up, what would I need to know about it that maybe I don't have in a quick glance?" [CS010]. They also want to let users know about known issues "so they don't have to dig through it themselves to figure it out" [CS010].

Curators anticipated other types of users and user questions. For example, one described tailoring datasets to a range of users: "Our data needs to be easy enough to use for the most novice user, but sophisticated for the more advanced user as well. And I think that that can happen by doing the details, making it easier" [CU017]. A curatorial supervisor considered how the dataset could be represented through crafting good metadata in supporting of use:

> And then even in our metadata. So I recently was talking to someone because I could tell them that the metadata ... Like the way they have worded was too internal-facing. I said, "That's not going to mean anything to users." If they don't understand, they're not going to hear about it. So we need to make it in a way that it's going to be something that makes sense to them. And is useful for them. [CS014]

And a third curator specifically linked the craft and subjectivity of curation with being able to conceive of how different users might perceive a dataset.

> I look at curation as, you know those technical aspects, there's do's and there's don't's, right and wrong. There's also some, I like to say artistry, subjectiveness do it and how I might perceive a group of people wanting to view the data. Another person might see it differently. [CU016]

The gestalt techniques curators used manifested differently in different cases. CU017 (above) expressed creativity in the information they chose to highlight for users. Several curators described differences in the order in which they approached curation tasks for a data study or in the time or level of detail they devoted to certain activities over others, such as developing the processing plan. This ability to assess the present data, conceive of the path to a future state, and conceive of a future representation is one aspect of craft exhibited by the curators.

4.2.2 Achieving standardization through judgement. Standards play a large role in data curation: at ICPSR, these include internal "house" standards set by ICPSR itself and external standards developed by the broader community, such as metadata standards or preservation best practices. However, the application of standards is far from rote. Curators use their expertise and make judgements about when, how, and why to apply standards throughout the curation process. Several participants said that because the data that ICPSR receives is just too diverse for strict standards to be feasible, curators must rely on their craft knowledge and expertise to navigate "gray areas" [CS012]. This can happen with "unusual datasets" in unique formats or with idiosyncratic structures [PM028], for particular archives with distinct user communities, or in instances where a principle investigator (PI) has requested what one participant called "a la carte" curation, where they do everything from one level, plus one task from another [CS009]. One supervisor said there are multiple workflows that stem from agreements with PIs, and therefore, one singular workflow isn't possible:

> I do think that we will always have more than one workflow just because sometimes the way proposals have to be written ... Sometimes project officers have a certain thing in mind. So, I think I mentioned earlier some PIs want a lot of involvement in disclosure review or the changes that we're making. So, for the demography archive, for pretty much most, if not all studies, we basically list out the examples of things that we're changing, and send it to the PI for approval. So, that's a different work flow than normally we just kind of can proceed. And then there's an archive within the criminal justice archive where the analysts want to use the data to do their analysis and then publish reports before we release the data. So, we have a process there were we do the first quality check then send it to them for review. They might send changes back to us, they might spend the next six months analyzing data. And so, we may not return to that study to make changes and/or do the second quality check and release it until months, or even a year later. [CS001]

Some curators expressed frustration with standards. Long-time curators in particular viewed the standards as living, malleable tools that change over time rather than unbreakable rules. Some went so far as to say that standards could be an obstacle to their work because they interfere with their preferred way of working. One curator felt that "some of the standards that we have get in the way of it when they're supposed to help it" [CU023]. A long-time curator observed:

> They're getting better. Previously, there was just a lot of unanswered questions in the document. A lot of ambiguity. … Sometimes I'm a little outspoken just because I've been around and... We call them standards. [CU017]

This curator went on to say:

> I don't necessarily agree that they're standards. They're somebody's opinion that got put down and then set as a standard. It's somebody's preference then they made it a standard, especially on sentence case for variable labels that one's like... Those aren't standards. That's somebody's preference and I don't like it, because I would just do it differently. Not that my way is right and they're wrong. It's just different ways of thinking. If I was the one creating that standard it'd be different because my preference and my thought process is different. [CU017]

Several curators further discussed the importance of focusing on user needs by applying standards creatively and flexibly. One curation supervisor acknowledged the importance of creativity to sidestep standards when the user was not served:

> So yeah, keeping those standards in mind, it's just sometimes you have to be creative. If there's something that you know users need to know who, put it in the summary field, don't put it in the collection notes, make it more visible. Maybe your tools don't allow us to make it as visible as we'd like. But there's always a way. [CS007]

Curators were also aware that applying the standards involved judgement calls and could be self-reflective on their comfort level in making some types of these calls:

> So there's a lot of judgement calls involved despite all of our efforts to write up standards and follow them, again it's human produced data and documentation. And there are still judgement calls to be made from grammar and spelling and capitalization to more serious matters. So there's a number of judgement calls that I feel comfortable making, such as those involving labels. But then on the other hand, if it's a really sort of hairy disclosure risk scenario, I would definitely check in with my supervisor on those. [CU024]

In deciding whether to apply standards or not, curators use expert judgement, creativity, and skill to achieve a standard outcome. One curation supervisor described their work as helping maintain the standards among curators but not setting or enforcing them [CS012]. Curators were also self-reflective about the standards and considered the data themselves as well as potential users when arriving at solutions that fell outside a "normal" application of the standards. Whether the curators were more respectful or skeptical of the standards, many discussed instances where the standards fell short of achieving a dataset fit for use.

4.2.3 Organizing curatorial actions. Though there is a *common* sequence of curatorial actions at ICPSR, there is no strict workflow; processing plans outline the work that's needed at a high level but not how it should be carried out. Curators use craft knowledge to sequence their technical work and to customize their work practices to the dataset at hand or their own preferences. As one supervisor described:

> We always say that curation isn't a linear process. There are a set of tasks that it makes sense to do in a particular order sometimes... I mean, we try to leave it up to the curator to what it works best for them because everyone has different ways of curating so... some people like to do metadata first, some people do that last, some people want to make all these data edits right away, some people want to focus on peripheral stuff. We try to get the processing plan done as soon as possible just because that helps expose all of the other issues that we might need to go to the project manager or the PI about. And that gives us more information about if we need to prioritize something particular. [CS013]

The curators themselves described considerable variability in how they ordered their work:

> I'm very collective, and it's not always the case, but just as I would prefer to be working on multiple curation projects at a time and instead of just focusing on one all day, every day until it's done, I like to have two or three going, if possible, just to break it up, break up my day, break up my focus. I also have that same approach for the tasks. So I might jump between different things. [CU016]

> Well the prioritization is to complete the plan and the disclosure risk worksheet. So that is where I start. And in completing those, I set the agenda, so to speak, for where it goes next. [CU024]

> I tend to start with the data, I tend to leave metadata to the end because I often find... it could go either way, right? You could do the metadata to inform how you approach the data, but I often find that going through the data, going through the questionnaire lets me fill in the metadata better. Yeah. So my first step is the plan and the worksheet, because that is the first part of the process, and there's checks involved with other people. And from there I usually tackle the data first. [CU024]

> I definitely jump around. [CU015]

The common thread throughout these different approaches is that curators draw on their own expertise to structure their work and days; they know what works best for them and how best to hone their attention for detailed, technical, and sometimes tedious work.

4.3 Coordination in service of curation

Above, we described how curators use craft practices to gain a gestalt understanding of a dataset's structure and then to organize their own work. This work is not done in a vacuum, however; curators must also coordinate with other stakeholders at ICPSR to proceed with this work and to clarify priorities. Because ICPSR's workflows resist standardization, curators and curation supervisors

must consult archive directors, project managers, data producers, and each other to ensure there is a consensus (if not agreement) on the best way to approach curating a study. This occurs throughout the curation process. For example, coordination occurs early on to determine where a study is placed in the curation queue and identify the level of curation it needs. Coordination can also occur later in the curation process if issues emerge requiring a decision about additional curation activities, which are required to make the study fit for use. Acts of communication are captured in the Jira ticket worklogs, but the content is often vague. Our interviews elucidated the frequency and critical place of coordination in the curation process. These include the following: prioritizing studies in the queue, specifying how data will be curated, and monitoring progress and alerts.

4.3.1 Prioritizing studies in the queue of deposits. Curation supervisors manage a large queue of studies waiting to be processed and assess how, when, and to whom to assign them based on the priorities and funding available to the various topical archives at ICPSR. This assessment includes tight coordination with archive directors, data producers, project sponsors, and project managers. Curation supervisors factor in a project's budget, promised deliverables, relevant external deadlines, and the potential impact of the study's release to determine placement in the queue. Two archive directors described this balance of considerations:

> Our funder really decides what to archive. I work with our project manager to [...] ensure that the curation team is prioritizing our data the way we want it prioritized. [...] And the project manager ensures that those [...] priorities get communicated to the curation team. [AD004]

> I would say that feedback from the funders influences both the curation levels and the priorities that we give to studies. So we do coordinate with our program officer. [...] If there are certain studies that are a high priority and that is something that we would then incorporate into Jira and into the curation requests so that we can adjust priorities and make sure that the highest priority work gets prioritized accordingly. [AD029]

Project managers also communicate priorities to the curation unit. They do this as a matter of routine through multiple reinforcing channels: project managers enter deadlines and rank relative priority in Jira tickets (e.g., Highest, High, Medium, Low), and they hold quarterly meetings with curation supervisors to "talk about the queue for a particular quarter" (PM025). However, as several curation and project management staff noted, priorities change, and the communication often involves significant back and forth:

> There's a lot of back and forth in terms of what their priorities are versus what we feel we can reasonably accomplish in a given time frame. And so sometimes that can get a little tricky. So in terms of like, if they say we have this and this, we're going to ask them which one is more important to them? I'm not going to try to figure that out, if I can assign them both I will, if I can't I'll make sure it's their highest priority. [CS010]

4.3.2 Negotiating levels of curation. Though the project managers initially define the work expected by choosing a curation level (1, 2, or 3), curators sometimes find that a given dataset needs more or different curation than originally planned. When this happens, they (and curation supervisors) must negotiate up and down the organizational chart to come to an agreement about how curation will proceed. Curators and curation supervisors negotiate with project managers, who coordinate with archive directors, who sometimes coordinate with PIs. A curation supervisor described the negotiation process that can be involved in curation level clarification:

> We have curation levels and the project manager reviews those levels and says, "Okay, I want this level of curation." And then we would review it to see if that's accurate. So that would be like a collaboration between the supervisor and the curator. So when

> [the curators] do their processing plan, they may identify, "Hey, they're asking me to do something that isn't in this level," or "they're asking me to do things but it should be like a level up or down." And then we also do a review of the plan and then we assess as well. [CS014]

We note here that the project managers may not necessarily get the same gestalt view of the data as the curators, so they trust the curators' views in further structuring work. As curation proceeds and curators get into the data, they often discover things that suggest several possible courses of action that can prompt discussion with curation supervisors, project managers, archive directors, and the data providers themselves. One curator described working with their supervisor to make final decisions. Though the curator ultimately defers to the curation supervisor, there's still a conversation about potential options:

> If we've identified an issue or something, [...] I might give [the supervisor] some options and then we talk about it a minute, and ultimately [...] I let her decide as the supervisor, especially when it comes to things on how to address confidentiality things. Those are definitely things that supervisors would like to have their approval on before things get out. It lessens the responsibility in a way on us by having that supervisor, or someone who's in charge, being able to make the final decisions [...] It's good to have, I think, other people's opinions on those things. Yeah, for my supervisor, it's definitely a conversation of, "Here's some possibilities of what we could do." [CU015]

Thus, there is some tension between respect for the curators' expertise and deep knowledge of their data, ICPSR's standards and decision-making hierarchy, and the overall budget for a project.

4.3.3 Monitoring progress via alerts, and navigating varying degrees of visibility. One mildly controversial method of facilitating coordination is the use of Jira tickets to monitor and record progress on a project. Curators, supervisors, and project managers communicate about the study via Jira ticket comments. Curators receive an alert every time tickets are updated, and the tickets act as a running log of the work performed on the study. Though project managers found Jira to be generally helpful, some curators characterized Jira as annoying or overwhelming. The deluge of alerts and documentation also made some curators feel micromanaged, as if Jira was keeping a running log of their work for their supervisors to review at any moment. As with any representation of work, however, the Jira tickets can be more or less precise. As one curator noted:

> So I have trouble... occasionally I have trouble keeping up with the ticket, the Jira ticket, where we're meant to tick off things as we go because I'm kind of doing a little bit of everything at once because every part has information you need that affects other parts. I often find myself quite close to the end and I'm like, "Oh shoot, I have to go update the ticket." [CU024]

The curators' feeling of sometimes being overly visible is mirrored by other concerns about curatorial work being underrecognized or that curatorial work is insufficiently visible. For instance, one project manager said that they felt curation skills were invisible to those who are more removed from it:

> I think that generally a lot of project managers and directors think that it's simple, simple syntax being applied but some of the challenges that come up while curating data can be quite complicated. It can take a certain level of skill. [PM026]

The centralization of the curation function has taken the curators out of the individual archives, and the implementation of a single, shared standard has altered their practice to align with the organization rather than with a single archive within ICPSR. In this new arrangement, ICPSR staff interact with curators at discrete points in the curation process, but few interact throughout the

process. Therefore, there is less opportunity to see how curators use complex representations to envision the data as fit for use and how they use judgement and creativity to achieve standardized outcomes for data. More coordination via intermediaries—whether Jira or project managers—becomes necessary to support curatorial work.

5 DISCUSSION

5.1 Understanding data curation and data archives

One of the motivations of this study was to gain a finer-grained understanding of work in a data archive, specifically focusing on curatorial actions. We developed a rich description of data curation at ICPSR—one that goes beyond procedural work with data and metadata, to include the expertise-driven decision-making involved in crafting data, and the coordination required to develop a consensus around curatorial priorities and activities. Our participants have shown that data curation is neither "magic" nor "janitorial" work [19, 58, 67] but rather is the result of technical skill enacted through craft practices. Indeed, we find that staff members in all roles bristle against characterizations of curation as something rote or mechanical. Curators do what what needs to be done to achieve the outcomes of a standard—even when not necessarily *following* a standardized workflow. This requires significant collaboration with other stakeholders in the data science workflow. Thus, the workflow is achieved but much of the actual work that made that happen disappears.

Our research makes two main contributions to understanding data curation, and thereby data, work. First, the description of "hands-on" technical tasks we provide in Section 4.1 expands an existing body of literature describing data curation practices in different contexts. Understanding different data (curation) practices is critical for building infrastructure, software tools, and ontologies that capture disciplinary contexts and for educating curators. For instance, Chao et al. [12] developed the Data Practices and Curation Vocabulary, which describes how a community (in that case, earth scientists) defines data curation. Comparison of our two frameworks reveals that ICPSR has much more detailed quality check protocols and that ICPSR's curators spend considerable time on tasks like "adding question text" that simply are not needed in the earth science fields. The diversity of curatorial actions shown in just these two papers highlights the need for further research into the specific curatorial workflows and communication regimes in different scholarly settings. It is well understood from research on data practices that there are significant domain differences in curation needs [2, 14, 23, 88]. Yet models of data curation rarely account for this diversity of practice, or provide guidance in how to navigate them.

Second, our work shows the vital role that craft practices play in successfully organizing curatorial work and applying and navigating standards. *In data work, we see craft manifesting as the ability to develop an abstract, gestalt representation of a data product and then envision how to make changes to that data product so that it is more fit for use. This work involves following best practices and creating a standardized product but not necessarily following a standardized workflow.* Furthermore, the kind of data curation carried out at ICPSR requires significant collaboration and consultation with other stakeholders. This extends prior work on craft in technical settings in the CSCW literature, most recently discussed by Muller et al. [53] in their summary of craft in the context of data science, as well as a more recent focus on craft in the LIS literature by Owens [58].

At ICPSR, we see clear alignments with some of Muller et al.'s account of craft in data work. Muller et al. [53] identified key characteristics of craft in CSCW, including "Conversation with materials: Through the conversation with materials, there is often a sense of intimacy with materials and media" and "Control: Craft-workers labor at an intersection of control and unpredictability." ICPSR curators repeatedly emphasized the importance of the "conversation with materials" in their work

through repeated descriptions of the contingency of their workflows and specific tasks. Likewise, ICPSR curators exist at the intersection of "control and unpredictability"—they are constrained into somewhat narrow roles by ICPSR's organizational structure, yet must navigate unpredictable and unique curation challenges for each dataset with stakeholders throughout and outside of ICPSR.

Our research further shows how craft practices "fit" into best practices and other standards for working with data; *in short, we have found that craft practices are necessary to enact best practices.* Data standards can vary in their application and results, based on variations in how they are enacted by a group [52]. Yet at ICPSR, we see *a standardized result arising from the nonstandardized application of standards via craft practices.* By giving curators the freedom to rely on their own skill to structure their work and make decisions, ICPSR is able to truly rely on them as the human-in-the-loop.

Accounting for the role of craft and expertise in data work is important in designing effective data workflows, training data workers, and better supporting data workers by showing the impacts of their work. Our work raises the following questions: 1) How do notions of "craft" complicate the development of data curation pipelines or the automation of data curation? We consider this question in Section 5.2. 2) How does understanding the craft involved in data work support data workers in gaining credit for their work and its impact? We address this question in Section 5.3. We close by considering implications for practice in Section 5.4.

5.2 Coordinating work in data curation: complicating "workflow" or "pipeline" views of data science and curation

One of our primary findings is that data curators at ICPSR must structure their own work within the context of their organization's structure and job descriptions and constraints. In this way, they and other stakeholders "work the workflow" and navigate across standards and up and down the organizational chart; they gain a gestalt view of not just the data at hand but also of the organization as a whole. Coordination and communication are key in this. In identifying coordination and craft practices as important parts of data curation work, we complicate not just technical accounts of data curation but also "workflow" or "pipeline" conceptions of data work. By "workflow" views, we mean conceptualizations of data curation as a sequential process, easily represented by a UML diagram or similar technique. These representations are quite common in CSCW and IS, where they are used to model curation processes at a high level [17, 25, 36, 40, 46, 94]), or to capture detailed change logs and provenance of a dataset [30, 31, 83, 96]. The models are common because of their utility; they represent complex processes in a way that is digestible, and they can act as boundary objects that help communication between disparate groups of stakeholders [21].

However, our work here underscores that data curation is more than the sum of its parts and that it involves much more than the objects that are curated; it is also a process in which distributed knowledge management decisions are made to facilitate information reuse [1]. Our research supports the notion that data curation is a highly collaborative process occurring across a distributed system over time. While some curation actions tend to occur in sequence, important components of curation work, like quality checks, are performed in parallel or iteratively throughout the curation process. Project-related communication is also embedded in all other curatorial actions, making it difficult to delineate. A closer look at project-related communication reveals the importance of discussion and delegation in curatorial work; for example, supervisors and curators often discuss how best to mitigate disclosive variables on a case-by-case basis, following risk minimization heuristics rather than hard rules. And though coordination strategies such as *prioritizing studies in the queue of deposits* and *specifying and clarifying how the data will be curated* may seem like they could fit neatly into a workflow diagram, in reality, they require a meta-level understanding of the curation workflow itself to proceed. Articulation work is needed to navigate a data science

workflow [57, 82], yet this labor can, somewhat ironically, be obscured in workflow-centric views. We want to be clear: we are not trying to discourage or dismiss workflow-based explorations of data work. Rather, we want to note the importance of continued, rich exploration of what goes on in and around each "box" of the diagram — lest we obscure that which we wish to reveal.

5.3 Revisiting the invisible nature of data curation

An additional impact of a workflow-centric view of data curation is that it can render individual curators as cogs in a machine—or, as Plantin [65] has described, factory workers on an assembly line. The alienation of assembly-line work leads curators to feel dissatisfied with their work and invisible. While we did not find that our curators felt the same levels of dissatisfaction or isolation in their work, we did find that curators—and even their managers, to a degree—described some concern that their work was not truly seen or appreciated. Invisibility can make it harder for these data workers to advance in their careers, lobby for salary increases, and participate fully in their fields [65]. We find that there are tensions, though, in making curatorial work totally visible. Below we discuss both the visibility and hypervisibility of curatorial work at ICPSR.

The craft and coordination in curatorial work at ICPSR are mostly invisible to data users. The public datasets hide the work that went into their creation precisely because they are standardized [64]. Even the documents that emerge from curation hide aspects of this work. While the Jira tickets contain descriptions of the high-level tasks, they do not provide a full account of the curators' labors and decision-making process. The existence of data curation standards makes the work seem routine even though all our interviewees recognize, to varying degrees, that curation requires technical skill, flexibility, and coordination.

At the same time, some aspects of curators' work is hypervisible within ICPSR through Jira tickets and other documentation. The Jira ticket worklogs and comments, especially, serve first to coordinate work and then to document it. And while Jira can document their labor and decisions, making their work visible, it can also open curators to negative side effects, such as micromanagement. For instance, Jira tickets make it possible for more powerful colleagues (e.g., archive directors) to monitor curators' work. As Suchman [77] and Yates [93] pointed out years ago, technologies that help workers coordinate locally can become mechanisms of global control by enabling surveillance and proscription. Thus, not all invisible work should be made visible. Bishop [8] uses Weber's concepts of "status contract" and "instrumental contract" to understand the changing relationships between employers and employees. She notes that status contracts—those that are about our relations to one another rather than our performance—often rely on the trust that results from these relationships and not from formal articulations of the work. At ICPSR, we see evidence of this status contract; by and large, those higher in the organization respect the skill and expertise of their curators. There is an understanding that some aspect of data work will always be invisible. The use of Jira tickets to monitor, however, threatens to replace this status contract with an instrumental one, in which the worker is valued for visible products.

How does viewing curation as a craft impact this (in)visibility? When supervisors, project managers, and archive directors view and treat curation as a craft, it supports the status contract between curators and higher management. It appreciates this data work as skilled labor and thereby "affords identity, status and a sense of connection to others in the enterprise and to the enterprise itself" (Nardi and Engeström [55] citing Bishop [8]). When we, as data practices researchers, data science educators, and CSCW theorists, argue for curation as craft, we too contribute to the support of this status contract. Thus, recognizing curation as craft is important to supporting labor arrangements that do not render the worker invisible even when the work is.

One impact of the invisibility of curation work is that outsiders underestimate its costs and value, and, by implication, the value of curators. Work like curation that is conducted in the background

is often taken for granted. Recent efforts to surface curatorial contributions to scholarship via structured metadata [80] or improvement of legacy data records [7] echo prior efforts, such as the Nursing Interventions Classification, to make work visible in efforts to legitimize both the work and workers [11, 74]. Here, we are pushing to recognize the labor needed to organize, understand, and negotiate the tidy boxes on workflow diagrams—and to recognize this work's seeming ineffability is important to preserve and respect.

5.4 Implications for practice

Better articulating the work and craft of data curation has several implications for practice. First and foremost, understanding the complex role that different forms of visibility play in data work may help us design technologies for users that move beyond reporting and surveillance [77]. As we described, many ICPSR curators bristled at the constant use of Jira because it made them feel hypervisible, monitored, and mildly harangued. We consequently ask: given a view of curation as craft rather than rote mechanical labor, what changes might we imagine for ticketing systems like Jira—ones that might lead the management system to serve the curators as well as their supervisors and managers?

Our work also has several implications for data curation training and education. Within the information sciences, considerable effort has been put into designing data curation curriculum for budding information professionals; much of this has been highly focused on articulating different versions of data curation workflows and describing data practices in different fields. The range of high-level curatorial actions we identified in Section 4.1 contributes to this tradition.

However, the greater contribution of our work is the importance of training data curators as craftspeople and not just technicians. As we quoted from Owens [58] previously, a craftful approach to curation is one that stays engaged with the "unresolved" and contingent aspects of curatorial work and that sees the "inherent messiness" of data work as a feature rather than a bug. Our work helps more specifically identify the strategies ICPSR curators use to navigate this unresolved messiness, particularly in how they use gestalt approaches to see the dataset as more than the sum of its parts. Though more work would be needed to better understand this process, we believe it is a promising direction for further curriculum develop—whether in data curation classes at the master's level or online lessons in the vein of Data Carpentry.[4]

One of the limitations of this study is ICPSR's unusual size and scope; they simply have a much larger and more well-organized data curation team than many other peer institutions and archives. It is possible that lessons learned here will not translate well to smaller contexts or teams. However, we believe the view of data work as being grounded in craft practices could be important to explore elsewhere. How do craft practices differ in smaller organizations or in teams where there are no dedicated curators? For those that think of data curation as a "wrangling" or "munging" process, how could adopting a craft perspective help guide their work and make it more reproducible?

6 CONCLUSION

Data curation is a critical component of data science and an important aspect of data work. Obscuring the work of data curation not only renders the labor and contributions of the data curators invisible but also makes it harder to tease out the impact curators' work has on the later usability, reliability, and reproducibility of data. In this paper we have made curatorial work visible through a case study of data curation at ICPSR, a large social science data repository. We have contributed a rich description of curatorial work at this site, including a range of technical curatorial actions, and the craft and coordination needed to successfully enact those actions. We

[4]https://datacarpentry.org

echo prior work calling for a craftful view of work with data: curation requires not just a rote following of standards and protocols but rather a creative, ongoing conversation with the data, with one's colleagues, and with one's community. Our work complicates workflow-based views of data curation, in that we find ICPSR curators do considerable work that can't be easily visualized with a UML diagram, and indeed, rely on craft practices to enable their workflow. We also find that ICPSR curators sit at an intersection between visibility and invisibility: their work is highly documented (and even monitored, to a degree), yet when they do their jobs well, it is invisible. Finding ways of selectively making curatorial work visible in service of curators will be key in supporting their work and professional development, as well as the development of data curation tools.

ACKNOWLEDGMENTS

Author contributions are as follows: **Andrea Thomer**: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Methodology, Formal Analysis, Supervision, Funding Acquisition; **Dharma Akmon**: Writing - Original Draft, Methodology, Formal Analysis, Funding Acquisition; **Jeremy York**: Writing - Review & Editing, Formal Analysis; **Allison R. B. Tyler**: Methodology, Investigation, Formal Analysis; **Faye Polasek**: Writing - Review & Editing, Formal Analysis; **Sara Lafia**: Writing - Original Draft, Writing - Review & Editing; **Libby Hemphill**: Conceptualization, Writing - Review & Editing, Supervision, Funding Acquisition, Project Administration; **Elizabeth Yakel**: Methodology, Investigation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding Acquisition.

We thank David Bleckley and Elizabeth Moss from ICPSR for their comments on earlier drafts. This material is based upon work supported by the National Science Foundation under grant 1930645. This project was made possible in part by the Institute of Museum and Library Services LG-37-19-0134-19.

REFERENCES

[1] Mark S. Ackerman and Christine Halverson. 1999. Organizational Memory: Processes, Boundary Objects, and Trajectories. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. Abstracts and CD-ROM of Full Papers (HICSS-32)*.

[2] Katherine G. Akers and Jennifer Doty. 2013. Disciplinary Differences in Faculty Research Data Management Practices and Perspectives. *Int. J. Digit. Curation* 8, 2 (2013), 5–26.

[3] Karen Baker and Florence Millerand. 2007. Articulation Work Supporting Information Infrastructure Design: Coordination, Categorization, and Assessment in Practice. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)* (Waikoloa, HI, USA). IEEE, Piscataway, NJ, USA, 242a–242a. https://doi.org/10.1109/HICSS.2007.88

[4] Alex Ball. 2012. *Review of Data Management Lifecycle Models*. University of Bath. http://opus.bath.ac.uk/28587/

[5] Stephen R. Barley and Julian E. Orr. 1997. Introduction: The Neglected Workforce. In *Between Craft and Science*. Cornell University Press, Ithaca, NY, USA, 1–20.

[6] William C. Barley, Jeffrey W. Treem, and Paul M. Leonardi. 2020. Experts at Coordination: Examining the Performance, Production, and Value of Process Expertise. *J. Commun.* 70, 1 (2020), 60–89.

[7] Bionomia n.d.. Bionomia. Retrieved July 14, 2021 from https://bionomia.net/

[8] Libby Bishop. 1999. Visible and Invisible Work: The Emerging Post-Industrial Employment Relation. *Comput. Support. Coop. Work* 8, 1-2 (March 1999), 115–126.

[9] Christine L. Borgman. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press, Cambridge, MA, USA.

[10] Christine L. Borgman, Andrea Scharnhorst, and Milena S. Golshan. 2019. Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse. *J. Assoc. Inf. Sci. Technol.* 70, 8 (Aug. 2019), 888–904.

[11] Geoffrey C. Bowker, Stefan Timmermans, and Susan Leigh Star. 1996. Infrastructure and Organizational Transformation: Classifying Nurses' Work. In *Information Technology and Changes in Organizational Work (IFIP Advances in Information and Communication Technology)*, Wanda J. Orlikowski, Geoff Walsham, Matthew R. Jones, and Janice I. Degross (Eds.). Springer US, Boston, MA, USA, 344–370.

[12] Tiffany C. Chao, Melissa H. Cragin, and Carole L. Palmer. 2015. Data Practices and Curation Vocabulary (DPCVocab): An Empirically Derived Framework of Scientific Data Practices and Curatorial Processes: Data Practices and Curation

Vocabulary (DPCVocab). *J. Assoc. Inf. Sci. Technol.* 66, 3 (March 2015), 616–633.
[13] Amy Cheatle and Steven J. Jackson. 2014. Digital Entanglements: Craft, Computation and Collaboration in Fine Art Furniture Production. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada). ACM, New York, NY, USA, 11. https://doi.org/10.1145/2675133.2675291
[14] Melissa H. Cragin, Carole L. Palmer, Jacob R. Carlson, and Michael Witt. 2010. Data Sharing, Small Science and Institutional Repositories. *Phil. Trans. R. Soc. A.* 368, 1926 (Sept. 2010), 4023–4038.
[15] Roderic Crooks and Morgan E. Currie. 2021. Numbers Will Not Save Us: Agonistic Data Practices. *Inf. Soc.* 37, 4 (2021). https://doi.org/10.1080/01972243.2021.1920081
[16] Peter T. Darch, Ashley E. Sands, Christine L. Borgman, and Milena S. Golshan. 2020. Library Cultures of Data Curation: Adventures in Astronomy. *J. Assoc. Inf. Sci. Tech.* 71, 12 (2020), 1470–1483.
[17] DataONE. 2015. Data Life Cycle. Retrieved July 14, 2021 from https://old.dataone.org/data-life-cycle
[18] Christine Dearnley. 2005. A Reflection on the Use of Semi-structured Interviews. *Nurse Res.* 13, 1 (2005), 19–28.
[19] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* MIT Press, Cambridge, MA, USA.
[20] Peter Dormer (Ed.). 1997. *The Culture of Craft: Status and Future.* Manchester University Press, Manchester, UK ; New York, NY.
[21] Paul Dourish. 2001. Process Descriptions as Organisational Accounting Devices: The Dual Use of Workflow Technologies. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work* (Boulder, CO, USA). ACM, New York, NY, USA, 52–60.
[22] Ingrid Erickson and Mohammad Hossein Jarrahi. 2016. Infrastructuring and the Challenge of Dynamic Seams in Mobile Knowledge Work. In *Proceedings of the 19th ACM conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, CA, USA). ACM, New York, NY, USA, 1323–1336.
[23] Ixchel M. Faniel, Rebecca D. Frank, and Elizabeth Yakel. 2019. Context from the Data Reuser's Point of View. *J. Doc.* 75, 6 (2019), 1274–1297.
[24] Ixchel M. Faniel and Ann Zimmerman. 2011. Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. *Int. J. Digit. Curation* 6, 1 (March 2011), 58–69.
[25] John L. Faundeen, Thomas E. Burley, Jennifer Carlino, David L. Govoni, Heather S. Henkel, Sally Holl, Vivian B. Hutchison, Elizabeth Martín, Ellyn T. Montgomery, Cassandra C Ladino, Steven Tessler, and Lisa S. Zolly. 2013. *The United States Geological Survey Science Data Lifecycle Model.* Technical Report. Reston, VA, USA. Open-File Report 2013-1265.
[26] Sebastian S. Feger, Paweł W. Wozniak, Lars Lischke, and Albrecht Schmidt. 2020. 'Yes, I Comply!': Motivations and Practices around Research Data Management and Reuse across Scientific Fields. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 141 (Oct. 2020), 26 pages. https://doi.org/10.1145/3415212
[27] Melanie Feinberg. 2017. A Design Perspective on Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, CO, USA). ACM, New York, NY, USA, 2952–2963. https://doi.org/10.1145/3025453.3025837
[28] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723
[29] Elihu M. Gerson and Susan Leigh Star. 1986. Analyzing Due Process in the Workplace. *ACM Trans. Inf. Syst.* 4, 3 (1986), 257–270. https://doi.org/10.1145/214427.214431
[30] Carole Goble, Robert Stevens, Duncan Hull, Katy Wolstencroft, and Rodrigo Lopez. 2008. Data curation + process curation=data integration + science. *Brief. Bioinform.* 9, 6 (Nov. 2008), 506–517.
[31] Carole A. Goble, Jiten Bhagat, Sergejs Aleksejevs, Don Cruickshank, Danius Michaelides, David Newman, Mark Borkum, Sean Bechhofer, Marco Roos, Peter Li, and David De Roure. 2010. myExperiment: A Repository and Social Network for the Sharing of Bioinformatics Workflows. *Nucleic Acids Res.* 38, Issue suppl 2 (July 2010), W677–W682. https://doi.org/10.1093/nar/gkq429
[32] Ann G. Green and Myron P. Gutmann. 2007. Building Partnerships among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives. *OCLC Systems & Services: Int. Digit. Libr. Perspect.* 23, 1 (Feb. 2007), 35–53. http://hdl.handle.net/2027.42/41214
[33] Libby Hemphill, Margaret L. Hedstrom, and Susan Hautaniemi Leonard. 2021. Saving Social Media Data: Understanding Data Management Practices among Social Media Researchers and Their Implications for Archives. *J. Assoc. Inf. Sci. Technol.* 72, 1 (Jan. 2021), 97–109. https://doi.org/10.1002/asi.24368
[34] Sharlene N. Hesse-Biber and Patricia Leavy. 2005. *The Practice of Qualitative Research* (third ed.). SAGE Publications, Los Angeles, CA, USA.
[35] Tony Hey, Stewart Tansley, and Kristin Tolle (Eds.). 2009. *The Fourth Paradigm: Data-intensive Scientific Discovery.* Microsoft Research, Redmond, WA, USA.
[36] Sarah Higgins. 2008. The DCC Curation Lifecycle Model. *Int. J. Digit. Curation* 3, 1 (Dec. 2008), 134–140.
[37] Caihong Huang, Jian-Sin Lee, and Carole L. Palmer. 2020. DCC Curation Lifecycle Model 2.0: Literature Review and Comparative Analysis. https://digital.lib.washington.edu:443/researchworks/handle/1773/45392

[38] ICPSR. 2020. ICPSR Curation Levels. https://www.icpsr.umich.edu/files/datamanagement/icpsr-curation-levels.pdf
[39] Eun Seo Jo and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. ACM, New York, NY, USA, 306–316.
[40] Lisa R. Johnston. 2014. *A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot.* Technical Report. Twin Cities, MN, USA.
[41] Lisa R. Johnston, Jacob Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, and Claire Stewart. 2018. How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries. *J. Librariansh. Schol. Commun.* 6, 1 (2018), eP2198. https://doi.org/10.7710/2162-3309.2198
[42] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, New York, NY, USA, 3363–3372.
[43] Helena Karasti and Karen S. Baker. 2004. Infrastructuring for the Long-term: Ecological Information Management. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004* (Big Island, HI, USA).
[44] Helena Karasti, Karen S. Baker, and Eija Halkola. 2006. Enriching the Notion of Data Curation in E-science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. *Comput. Support. Coop. Work* 15, 4 (Oct. 2006), 321–358. https://doi.org/10.1007/s10606-006-9023-2
[45] Karina Kervin, Robert B. Cook, and William K. Michener. 2014. The Backstage Work of Data Sharing. In *Proceedings of the 18th International Conference on Supporting Group Work* (Sanibel Island, FL, USA) *(GROUP '14)*. ACM, New York, NY, USA, 152–156.
[46] Sean Kross and Philip J. Guo. 2021. Orienting, Framing, Bridging, Magic, and Counseling: How Data Scientists Navigate the Outer Loop of Client Collaborations in Industry and Academia. (2021). https://doi.org/10.48550/arXiv.2105.05849 arXiv:2105.05849
[47] Sara Lafia, Andrea Thomer, David Bleckley, Dharma Akmon, and Libby Hemphill. 2021. Leveraging Machine Learning to Detect Data Curation Activities. In *2021 IEEE 17th International Conference on eScience (eScience)* (Innsbruck, Austria). IEEE Computer Society, Los Alamitos, CA, USA, 149–158. https://doi.org/10.1109/eScience51609.2021.00025
[48] Margaret D. LeCompte and Jean J. Schensul. 2012. *Analysis and Interpretation of Ethnographic Data: A Mixed Methods Approach* (second ed.). Rowman & Littlefield, Lanham, MD, USA.
[49] Helena M. Mentis, Ahmed Rahim, and Pierre Theodore. 2016. Crafting the Image in Surgical Telemedicine. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, CA, USA) *(CSCW '16)*. ACM, New York, NY, USA, 744–755.
[50] Merriam-Webster. n.d.. Craft Definition & Meaning. https://www.merriam-webster.com/dictionary/craft
[51] Matthew B. Miles, A. Michael Huberman, and Johnny Saldaña. 2014. *Qualitative Data Analysis: A Methods Sourcebook* (third ed.). SAGE Publications, Thousand Oaks, CA, USA.
[52] Florence Millerand and Geoffrey C. Bowker. 2009. Trajectories and Enactment in the Life of an Ontology. In *Standards and Their Stories.*, Susan Leigh Star and Martha Lampland (Eds.). Cornell University Press, Ithaca, NY, USA, 149–165.
[53] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300356
[54] Tahani Nadim. 2016. Data Labours: How the Sequence Databases GenBank and EMBL-bank Make Data. *Sci. Cult.* 25, 4 (Oct. 2016), 496–519. https://doi.org/10.1080/09505431.2016.1189894
[55] Bonnie A Nardi and Yrjö Engeström. 1999. A Web on the Wind: The Structure of Invisible Work. *Comput. Support. Coop. Work* 8, 1 (March 1999), 1–8. https://doi.org/10.1023/A:1008694621289
[56] National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Committee on Science, Engineering, Medicine, and Public Policy, Board on Research Data and Information, Division on Engineering and Physical Sciences, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Analytics, Division on Earth and Life Studies, Nuclear and Radiation Studies Board, Division of Behavioral and Social Sciences and Education, Committee on National Statistics, Board on Behavioral, Cognitive, and Sensory Sciences, and Committee on Reproducibility and Replicability in Science. 2019. *Reproducibility and Replicability in Science.* National Academies Press, Washington D.C., USA.
[57] Andrew B. Neang, Will Sutherland, Michael W. Beach, and Charlotte P. Lee. 2021. Data Integration as Coordination: The Articulation of Data Work in an Ocean Science Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3 (2021), 1–25. https://doi.org/10.1145/3432955
[58] Trevor Owens. 2018. *The Theory and Craft of Digital Preservation.* Johns Hopkins University Press, Baltimore, MD, USA.

[59] Carole L. Palmer. 2006. Weak Information Work and "Doable" Problems in Interdisciplinary Science. *Proc. Am. Soc. Inf. Sci. Technol.* 43, 1 (2006), 1–16. https://doi.org/10.1002/meet.14504301108

[60] Carole L. Palmer, Melissa H. Cragin, and Timothy P. Hogan. 2007. Weak Information Work in Scientific Discovery. *Inf. Process. Manage.* 43, 3 (2007), 808–820. https://doi.org/10.1016/j.ipm.2006.06.003

[61] Carole L. Palmer, Nicholas M. Weber, Trevor Muñoz, and Allen H. Renear. 2013. Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data. *Arch. J.* (2013).

[62] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and Its (dis)contents: A survey of Dataset Development and Use in Machine Learning Research. *Patterns* 2, 11 (Nov. 2021). https://doi.org/10.1016/j.patter.2021.100336

[63] Kathleen H. Pine, Claus Bossen, Yunan Chen, Gunnar Ellingsen, Miria Grisot, Melissa Mazmanian, and Naja Holten Møller. 2018. Data Work in Healthcare: Challenges for Patients, Clinicians and Administrators. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18)*. ACM, New York, NY, USA, 433–439. https://doi.org/10.1145/3272973.3273017

[64] Jean-Christophe Plantin. 2019. Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science. *Sci. Technol. Human Values* 44, 1 (Jan. 2019), 52–73. https://doi.org/10.1177/0162243918781268

[65] Jean-Christophe Plantin. 2021. The Data Archive as Factory: Alienation and Resistance of Data Processors. *Big Data & Soc.* 8, 1 (2021), 205395172110075. https://doi.org/10.1177/20539517211007510

[66] Line Pouchard. 2016. Revisiting the Data Lifecycle with Big Data Curation. *International Journal of Digital Curation* 10, 2 (May 2016), 176–192.

[67] Katie Rawson and Trevor Muñoz. 2019. Against Cleaning. In *Debates in the Digital Humanities 2019*, Matthew K. Gold and Lauren F. Klein (Eds.). University of Minnesota Press, 279–292. https://doi.org/10.5749/j.ctvg251hk.26

[68] Daniela K. Rosner, Samantha Shorey, Brock R. Craft, and Helen Remick. [n.d.]. Making Core Memory: Design Inquiry into Gendered Legacies of Engineering and Craftwork. In *Proc. ACM on Hum.-Comput. Interact.* ACM, New York, NY, USA, 1–13.

[69] Herbert J. Rubin and Irene S. Rubin. 2012. *Qualitative Interviewing: The Art of Hearing Data* (3rd ed.). SAGE Publications, Thousand Oaks, CA, USA.

[70] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. 2021. "Everyone Wants to do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan). ACM, New York, NY, USA. https://doi.org/10.1145/3411764.3445518

[71] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 1–37. https://doi.org/10.1145/3476058

[72] William A. Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opin. Q.* 19, 3 (1955), 321–325. https://doi.org/10.1086/266577

[73] Richard Sennett. 2008. *The Craftsman.* Yale University Press, New Haven, CT, USA.

[74] Susan Leigh Star and Anselm Strauss. 1999. Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Comput. Supp. Coop. Work* 8, 1 (March 1999), 9–30. https://doi.org/10.1023/A:1008651105359

[75] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Avignon, France). Association for Computational Linguistics, Stroudsburg, PA, USA, 102–107.

[76] Besiki Stvilia, Charles C. Hinnant, Shuheng Wu, Adam Worrall, Dong Joon Lee, Kathleen Burnett, Gary Burnett, Michelle M. Kazmer, and Paul F. Marty. 2013. Studying the Data Practices of a Scientific Community. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (Indianapolis, IN, USA) *(JCDL '13)*. ACM, New York, NY, USA, 425–426. https://doi.org/10.1145/2467696.2467781

[77] Lucy Suchman. 1995. Making Work Visible. *Commun. ACM* 38, 9 (Sept. 1995), 56–64. https://doi.org/10.1145/223248.223263

[78] Alex S. Taylor, Siân Lindley, Tim Regan, David Sweeney, Vasillis Vlachokyriakos, Lillie Grainger, and Jessica Lingel. 2015. Data-in-Place: Thinking through the Relations between Data and Community. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea). ACM, New York, NY, USA, 2863–2872. https://doi.org/10.1145/2702123.2702558

[79] Carol Tenopir, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PloS ONE* 10, 8 (2015), e0134826. https://doi.org/10.1371/journal.pone.0134826

[80] Anne E. Thessen, Matt Woodburn, Dimitrios Koureas, Deborah Paul, Michael Conlon, David P. Shorthouse, and Sarah Ramdeen. 2019. Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations

of the RDA/TDWG Working Group. *Data Sci. J.* 18, 1 (Nov. 2019), 54. https://doi.org/10.5334/dsj-2019-054
[81] Andrea K. Thomer. 2022. Integrative Data Reuse at Scientifically Significant Sites: Case Studies at Yellowstone National Park and the La Brea Tar Pits. *J. Assoc. Inf. Sci. Technol.* 73, 3 (2022), 1155–1170. https://doi.org/10.1002/asi.24620
[82] Andrea K. Thomer, Michael Bernard Twidale, and Matthew J. Yoder. 2018. Transforming Taxonomic Interfaces: "Arm's Length" Cooperative Work and the Maintenance of a Long-lived Classification System. *Proc. ACM on Hum.-Comput. Interact.* 2, CSCW (2018), 1–23. https://doi.org/10.1145/3274442
[83] Andrea K. Thomer, Karen M. Wickett, Karen S. Baker, Bruce W. Fouke, and Carole L. Palmer. 2018. Documenting Provenance in Noncomputational Workflows: Research Process Models Based on Geobiology Fieldwork in Yellowstone National Park. *J. Assoc. Inf. Sci. Technol.* 69, 10 (2018), 1234–1245. https://doi.org/10.1002/asi.24039
[84] Mary Vardigan, Pascal Heus, and Wendy Thomas. 2008. Data Documentation Initiative: Toward a Standard for the Social Sciences. *Int. J. Digit. Curation* 3, 1 (Dec. 2008), 107–113. https://doi.org/10.2218/ijdc.v3i1.45
[85] Jullian C. Wallis, Christine L. Borgman, Matthew S. Mayernik, and Alberto Pepe. 2008. Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research. *Int. J. Digit. Curation* 3, 1 (Dec. 2008), 114–126. https://doi.org/10.2218/ijdc.v3i1.46
[86] Hadley Wickham. 2014. Tidy Data. *J. Stat. Softw.* 59, 10 (2014), 1–23. https://doi.org/10.18637/jss.v059.i10
[87] Mark D. Wilkinson, Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* 3 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18
[88] Michael Witt, Jacob Carlson, D. Scott Brandt, and Melissa H. Cragin. 2009. Constructing Data Curation Profiles. *Int. J. Digit. Curation* 4, 3 (2009), 93–103. https://doi.org/10.2218/ijdc.v4i3.117
[89] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Wikipedia Talk Labels: Personal Attacks.
[90] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17, Vol. 11)*. ACM, New York, NY, USA, 1391–1399.
[91] Elizabeth Yakel. 2007. Digital Curation. *OCLC Systems & Services: Int. Digit. Libr. Perspect.* 23, 4 (Nov. 2007), 335–340. https://doi.org/10.1108/10650750710831466
[92] An Yan, Caihong Huang, Jian-Sin Lee, and Carole L. Palmer. 2020. Cross-disciplinary Data Practices in Earth System Science: Aligning Services with Reuse and Reproducibility Priorities. In *Proceedings of the Association for Information Science and Technology* (Virtual Conference), Vol. 57. Association for Information Science and Technology, Silver Spring, MD, USA, e218. https://doi.org/10.1002/pra2.218
[93] JoAnne Yates. 1989. *Control through Communication: The Rise of System in American Management.* Johns Hopkins University Press, Baltimore, MD, USA.
[94] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (May 2020), 1–23. https://doi.org/10.1145/3392826
[95] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. 2015. Beyond Frontal Faces: Improving Person Recognition Using Multiple Cues. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. cv-foundation.org, Boston, MA, USA, 4804–4813.
[96] Jun Zhao, Jose Manuel Gomez-Perez, Khalid Belhajjame, Graham Klyne, Esteban Garcia-Cuesta, Aleix Garrido, Kristina Hettne, Marco Roos, David De Roure, and Carole Goble. 2012. Why Workflows Break–Understanding and Combating Decay in Taverna Workflows. In *2012 IEEE 8th International Conference on e-Science* (Chicago, IL, USA). IEEE, Piscataway, NJ, USA, 1–9. https://doi.org/10.1109/eScience.2012.6404482
[97] Ann S. Zimmerman. 2008. New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Sci. Technol. Human Values* 33, 5 (2008), 631–652. https://doi.org/10.1177/0162243907306704

A INTERVIEW PROTOCOL

Background

- How long have you been a [ROLE]?
  - What's your academic background?

  - Do you have any prior experience as a data manager, curator, etc.
- I would like to start by hearing more about what your [ROLE] work at ICPSR.
- What is your role in the curation process?
  - Would you describe the chain of command? (e.g., Archive directors, curators)
  - How do you work together to make decisions?
    * How do you interact with project managers?
    * Do you feel involved in making judgement calls?
    * When a study seems like it falls between two levels of curation, who determines which level to assign it?
      · What factors are important to this determination?
  - Can you describe your overall workflow to me?
  - Is your work specialized to an archive/domain/data type?
- How much curation do the datasets you work with typically need?
  - Do they tend to arrive in the same state?
- Would you tell us a bit about the scholarly community that uses your archive?
- What type of relationship does the archive have with the scholarly community reusing its data?

Curation

- [If applicable] What types of interactions do you have with the curation unit?
- How involved are you in curation decisions?
  - When does this occur (grant proposal, initiation of a grant/project planning, before/during data sharing)?
  - [If applicable] Were you involved in the recent decision to make ICPSR curation workflows more systematic? If so, can you tell us what led to that decision?
  - How have recent changes to curation workflows at ICPSR changed your involvement in curation decisions, if at all?
  - Is there a formal process?
  - Are these decisions always easy to implement?
- Which curation activities add the most value to your archive/datasets?
  - Why do you say this?
- How do you prioritize different curatorial activities?
  - How do you know when a dataset is "done" being curated?
  - How involved are you in making judgement calls (e.g., between levels, between different curatorial actions)?
- Are the curation levels well defined?
  - Do you think they work for most studies?
    * Why or why not?
- How has your job changed since the curation reorganization? [Tailor to ROLE and BACKGROUND]
- Do your data reusers or designated community provide input into curatorial decisions?
- How has the curation provided by ICPSR changed the use or impact of your collections?
- Do you ever question the amount of curation planned for or being applied to a dataset?
- Is there additional/different curation you'd like to see applied to some of your datasets?
- What metrics would you propose or like to guide the level of curation of data?

Impact and metrics

- What type of impact would you like your archive to have?
  - How close is the archive to achieving this goal?

  - Where would you like to see the impact of your archive in 5 years?
- Are you aware of any metrics at ICPSR guiding the curation process?
- What metrics do you currently use to measure the impact of your collection, if any?
- What metrics would you propose to measure the impact of your collection?
- [If applicable] Do you plan or discuss your curation work with anyone else at ICPSR?
  - If yes, how do their comments impact your curatorial decisions?
- [If applicable] Do you consider the potential impact of the dataset during curation?
  - If so, how?
- [If applicable] How does your work add value to the datasets you curate?
  - Probe if not answering specifically: application of standards, metadata
    * Which ones?
- How do you see x (e.g., metadata, data cleaning, etc.) having impact on the datasets?
  - [If applicable] In what ways do the JIRA tickets document the curation work you've done?
  - [If applicable] In what ways do the JIRA tickets *not* document the curation work you've done?
  - [If applicable] Do you find JIRA intrusive?
- Is there a dataset that you've worked on that's had substantial impact?
  - If so, could you describe what it was and what impact it made?
  - What contribution did your curatorial work have on this dataset's impact?
- How would your designated community define impact of the collections?
- What impact would the faculty that contribute data to your archive want the collection to have?
  - Do you believe that data sharing and reuse should be considered for promotion and tenure?
  - How broadly is this shared in the scholarly community served by the archive?
- What kind of impact do you want your work to have? Your collections to have?

Reuse

- Do you consider data reusers during the curation process?
  - If so, what are the significant characteristics or properties (e.g., information about the data that is important for effective preservation management or reuse) you think are important to capture to enable data reuse?
- Do you interact with data reusers?
  - If yes, do their comments impact your curatorial decisions?
- What are the greatest barriers in reusing collections from your archive?
- What kinds of input or questions do you get from data reusers?
- What do reusers tell you about the value of different datasets?
- What kind of reuse would you like to facilitate in the future?

Wrap-up

- Do you have any questions for us, or about this project?

Received July 2021; revised November 2021; accepted April 2022