Recovering and Reusing Historical Data for Science: Retrospective Curation Practices Across Disciplines

Amanda H. Sorensen(✉), Camila Escobar-Vredevoogd, Travis L. Wagner, and Katrina Fenlon

University of Maryland, College Park, MD 20742, USA
asorens1@umd.edu

Abstract. While data curation research and practice have provided a growing body of guidance for and tools to support the curation, sharing, and reuse of recent and future scientific data, attention to retrospective data curation has been limited. The Recovering and Reusing Archival Data for Science project draws on semi-structured interviews with scientists and data curators to investigate data recovery and reuse efforts focused on historical data, or data drawn from legacy research materials, across a wide range of institutional, disciplinary, and research contexts. This paper describes selected findings related to (1) the perceived value of historical data for current and future scientific research; (2) challenges particular to recovering historical data; and (3) ethical quandaries that arise in historical data recovery and reuse. These findings shed light on the potential impact of historical data recovery and implications for retrospective data curation practices in support of active scientific research across disciplines.

Keywords: Scientific data · Data recovery and reuse · Legacy data · Data curation · Data sharing

1 Introduction

Masses of potentially useful scientific data are hiding in the unprocessed, accumulated collections of scientific research institutions, repositories, and archives—in historical research records, in the papers of retired scientists, on hard drives of data from concluded projects, in boxes of historical publications, in working files and field notes. These sources hold data that may be keys to advancing research in the sciences and beyond. Yet, these data and documents often remain hidden, at risk of being lost or destroyed by technical obsolescence or gradual obscurity. Despite increasing recognition within various scientific disciplines of the potential value of data in archival records or in historical research materials, research and practice in data curation have focused on saving current and future data for reuse. Increasing scientific research relying on legacy data highlights the need for study on primarily retrospective data curation, focused on recovering reusable data from the historical record or defunct research materials.

I. Sserwanga et al. (Eds.): iConference 2023, LNCS 13971, pp. 14–28, 2023.
https://doi.org/10.1007/978-3-031-28035-1_2

The "Recovering and Reusing Archival Data for Science" project (RRAD-S) addresses the question: What opportunities and challenges confront the recovery and reuse of historical or defunct data for active scientific research across disciplines and organizational contexts? The impetus for this project began in case studies of recovering useful data from historical data collections at the National Agricultural Library. Having developed tools to assist memory institutions with data rescue [1], our research turned to a systematic study of the wider landscape of historical data recovery, through a semi-structured interview study with scientists and data curators. This research builds upon prior work on recovering or rescuing data at risk of loss, including from the Research Data Alliance Data Rescue Interest Group and the CODATA Data-at-Risk Task Group [2, 3]. We aim to build upon their progress with an empirical study of scientific and curatorial practices across a range of disciplines, organizations, and research contexts.

This paper discusses selected outcomes of this research pertaining to (1) the perceived value of historical data for current and future scientific research; (2) challenges particular to recovering data from historical sources; and (3) ethical quandaries that arise in historical data recovery and reuse. Our findings shed light on practices that have been ongoing for decades across disciplines, both within specific scientific projects and in the daily professional work of archivists and curators, but which have remained largely invisible to the wider body of literature in data curation. These practices have long been obscured by our focus—within the domain of data curation research and practice—on current and future scientific data, as opposed to data from archival settings or from long-defunct or historical research projects. In addition, scientists from wide-ranging disciplines have undertaken data recovery efforts but do not always publish on their recovery and curation practices (preferring to publish, instead, on the scientific outcomes of their analyses of recovered data). Where they do publish about their recovery practices, their efforts tend to remain siloed within a single discipline despite the relevance of their approach and insights to recovery efforts in other domains. There remains, too, a disconnect between archivists' and other professional curators' work on data recovery—which often focuses on recovery to support open-ended reuse of data—and the data recovery work of scientists who are undertaking the effort to support research on a specific question.

2 Prior Work

Data recovery is the process of enabling the sustained use and reuse of data that would otherwise go unused [4]. What data recovery looks like in practice tends to vary widely in different contexts. In general, data targeted by recovery efforts is salvaged from digital or analog sources that have been compromised by time, technological decay, or the gradual creep of obscurity. This includes data that reside on defunct hardware or inaccessible storage media, websites or digital publications no longer being maintained, databases forgotten on unplugged hard drives, data tables lurking among the unsorted papers of retired scientists, unprocessed boxes of photographs in deep storage, or in spaces and platforms—in the cloud and on the ground—affected by natural disasters, war and conflict, political shifts, etc.

The term *data recovery* marks a distinctive area within the wider landscape of data curation, defined as the ongoing management of research data through its lifecycle of interest and usefulness [5]. Data curation broadly encompasses practices such as data appraisal, description, transformation, and preservation measures necessary to keep data useful over time [6, 7]. Within this landscape, data recovery emphasizes aspects of curation applied to data that are no longer in use: data at risk of being lost or corrupted, and data from the past. Mayernik et al. [8] offer a matrix for understanding the risk factors that compromise the availability and usefulness of research data, including (but not limited to) losses of funding, losses of contextual knowledge, catastrophes, changes in legal status or ownership, and cybersecurity breaches. The challenges facing the preservation of scientific data are numerous, even for "healthy" data currently embedded in preservation systems or surrounded by users, funding, and supportive tools. Historical or otherwise defunct data face the same challenges and more, compounded by the passage of time, divorce from their original contexts, the inaccessibility of data creators, and technological deterioration and obsolescence.

Prior work from major professional organizations focused on data curation, including the Research Data Alliance Data Rescue Interest Group and the CODATA Data-at-Risk Task Group, have illuminated the need for cross-sector collaboration to build networks of support for preserving historical scientific data and supporting its reuse across disciplines [2, 3]. Yet, much of the on-the-groundwork of data recovery done in the sciences remains disconnected from parallel work in other scientific disciplines, and from the professional domain of data curation.

The most well-known data recovery initiatives stem from large-scale "community science," "citizen science," or crowdsourcing projects. But the long tail of data recovery efforts is largely invisible, going unpublished and unfunded, often serving the localized purposes of a single project or lab. Many varieties of recovery projects are ongoing every day across scientific domains, including documented efforts in climatology, astronomy, geology, pharmacology, oceanography, agriculture, etc. These efforts may leverage the work of crowds of volunteers, of automated approaches, or may rely on the manual labor of solo curators. Some fields prefer the term "data rescue", but in information science that term tends to have a narrower denotation of distributed, grassroots, and politically-motivated efforts such as those of the "Data Refuge" initiative, a widespread effort to save climate change data from administrative turnover after the 2016 U.S. presidential election [9].

"Data rescue" provides an alternative framing for recovery. Like recovery, rescue highlights the abundance of data, scientific or otherwise, in need of retrieval from dire, curatorial circumstances. However, rescue also advocates for the reevaluation of existing data, the identification of unacknowledged data sets, and, generally speaking, a more communal and crowdsourced approach to data curation [9]. In the context of archival practice, this includes a focused attempt to identify data often overlooked within commercial, proprietary curation software, and data produced within rather than outside of archives [10]. Such work also requires deliberate and adaptive cross-institutional collaboration, much of which necessitates building proactive preservation plans into data creation, curation, and management [11]. Further, since data rescue work often occurs in response to larger systemic issues of data value, it tends to be inherently activist, responding to perceived threats relative to shifts in governmental administrations, funding, and public support for scientific endeavors [12].

Regardless of the specific impetus or disciplinary context, all data recovery efforts, at base, aim to enable the possibilities of new knowledge from extant evidence. While data recovery encompasses a potentially vast range of tools, techniques, and strategies, depending on the data and the context for recovery, most documented recovery projects entail two basic stages: identifying potentially useful data, and performing systematic conversions of data or sources into more sustainable, useful formats [13, 14]. These efforts are invariably resource-intensive [15, 16]. Recovery efforts may serve myriad specific research purposes, but all those purposes can be understood within the frame of *reuse*. Data recovery supports the reuse of scientific data to serve contemporary, ongoing, or future scientific research. Recovered data can, in parallel, support historical research and social studies of science. Data may be recovered in pursuit of a specific research objective, which is often the case when recovery is done by scientists, or to support open-ended possibilities of reuse, which tends to be the case when recovery is led by data curation professionals within knowledge or memory organizations, such as a libraries, archives, or data repositories. Historical data has been shown to support longitudinal and meta-analyses, computational modeling, and cross-disciplinary research in various domains. Pasquetto et al. [17] offer distinctions among different kinds of data reuse: reuse to serve the reproducibility or replication of scientific research; independent reuse, in which data are deployed to answer novel questions; and integrative reuse, in which data are combined with other data in order to serve comparisons, new models, or new research questions altogether.

Most of the rich literature of theory and practice on data sharing and reuse focuses on data from current and future science: on data sharing practices among scientists, on data repositories and open infrastructures to support data sharing, and on standards, practices, and tools that allow curation professionals and scientists themselves to capture adequate contextual information about data to support reuse [17–19]. This literature has focused largely on the increasingly professionalized roles of data curators, and on scientific practices at their intersections with institutions like repositories.

In contrast, data recovery is distinguished by focusing on data that may never have been shared as such. It focuses on data that predate or have otherwise slipped through the growing infrastructures and best practices supporting open science. These data were not necessarily created with open-ended futures of broad access and reuse in mind. As a result, they exist in forms that are not readily accessible, whether by people or machines. Such data tend to arise from grassroots efforts, both within and independent of curation institutions. Relative to data curation more broadly, data recovery is poorly studied. There is a distinct need for cross-disciplinary, empirical research on facilitating the reuse of data that are not already amenable to use *as data*.

In addition, we need to understand the potential ethical hazards and sociotechnical implications of data recovery and reuse in different contexts. Like the shift from data recovery to data rescue, questions of ethical data reuse must navigate complexities around consent and beneficence, both for relevant communities of origin and original data curators. As a clear-cut example, data in health sciences often raise questions around benefits to both individual patients and broader advances within medicine [20]. In addition, shifts towards transparent models of data reuse and economic benefit continue to emerge in response to discussions both within academic and popular spaces on the use of data from historically marginalized populations, through practices such as biobanking [21]. Beyond health information—and the privacy and exploitation risks so visible in that context—documented ethical concerns in data recovery and reuse pertain to data sovereignty, community ownership and beneficence, creator intent, and ethics of access, e.g. [22, 23].

Ethical data reuse also necessarily factors in the role that scientific data transparency plays in the advancement of global knowledge [24]. Of course, this framing ignores broader sociocultural issues latent in the open sharing of scientific data, most prominently questions tied to data ownership and authorship within academic publication settings [24]. Scientific competition and ideologies around citation metrics and mantras of "publish or perish" dissuade scientists from sharing data out of legitimate fears of poaching [25]. To alleviate some of these concerns, models for identifying otherwise defunct data imagine new venues of data reuse, colloquially referred to as "data thrifting" [26]. This project offers a start on addressing these gaps in our knowledge.

3 Methods

This study comprised 23 semi-structured interviews with practitioners engaged in recovering and reusing historical, scientific data in a wide range of disciplinary, organizational, and research contexts. Our interview participants included research scientists, science librarians, curators, archivists, and volunteers working with crowdsourcing platforms, digital humanities centers, museum collections, scientific libraries, universities, academic organizations, and within many other contexts. Some of these participants are professionally trained data curators with academic backgrounds in library and information science. However, other participants, including research scientists, were not formally trained in digital curation methods.

The goal of interviews was to capture a broad range of data curation practices specific to historical data recovery and reuse. This phase of the research builds upon prior case studies of historical data collections at the United States Department of Agriculture's National Agricultural Library (NAL). In these forerunning case studies, reported in Shiue et al. [1], the research team undertook the curation of historical data from three diverse NAL special collections, including analog data, such as handwritten field notes, and tabular data on paper from early 20th-century collections of high scientific impact, and born-digital data from the donated papers of a recently retired scientist, much of which existed in obsolete and proprietary file formats. As we observed the myriad challenges that arose in these original case studies of data recovery and reuse and encountered the various communities in different sub-disciplines of agriculture doing related work largely without sharing their processes or outcomes, we identified a need for a broader overview of the landscape of historical data recovery and curation work across scientific fields, and the range of people, roles, and approaches involved.

Our interviews broached questions about participants' objectives for and experiences with scientific data recovery and reuse, how they identify data worthy of recovery, how they went about data recovery, and what challenges they encountered. Interviews took roughly one hour to complete and were audio-recorded with the permission of participants. All audio recordings were transcribed using Otter.ai and recordings were deleted upon completion of transcription. The interviews were then subject to qualitative content analysis.

The research team collaboratively built a codebook emically focusing on themes and key concepts as they emerged from participant discussions [27]. The codebook included 15 codes related to themes, including institutional practices, policies, curation practices, and data value to name a few. For the purposes of this paper, the team randomly selected a sample of 10 interviews to begin preliminary analysis. We coded these transcripts with qualitative coding software *NVivo*. To assure validity and intercoder reliability, the research team engaged in the constant comparative method and discussed discrepancies in coding [28]. When emergent codes or themes arose, the team would discuss new codes and reapply coding as necessary. Table 1 below represents a sample of codes, their definitions, and relevant quotes from participants. When writing, interview quotations were selected based on how they summarize key themes and perspectives succinctly. To assure participant anonymity, potentially identifying information in some quotations has been removed and replaced with a relevant descriptor, given in square brackets. Each interview participant has also been assigned an identifier (e.g., “P01”), which serves as their pseudonym to ensure their anonymity, as promised in our participation consent forms.

Table 1. Sample of codebook used in analysis

Code	Brief definition	Sample quote(s)
Data sharing challenges	Specific obstacles or barriers to data sharing, data exchange, or data transfer in any setting – whether in an institutional setting, between users and repositories, etc.	P01: “researchers hate sharing their data, they hate the public access policy, because it requires new work from them”
Formats	Particular formats, file formats, documentation practices or policies, metadata standards (formal or informal), or other facets of representation and description	P01: “Yeah, we have Microsoft Excel spreadsheets. We have CSVs, we have database files, some created in Access, some created in SQL, some created in other languages”
Evaluation	How participants gauge success or completion, how they evaluate their outcomes, indicators of success, completion, or impact	P02: “To make it more reproducible and more useful to more people if I finally get all the data extracted from these journals, these conference proceedings that I’ve set out for myself, that would be good”

4 Preliminary Findings

This research surfaced numerous challenges and opportunities for historical data recovery and reuse. Many of the findings parallel familiar challenges from data curation more broadly, including the labor-intensiveness of preparing data for reuse, whether of migrating data to new formats or documenting them sufficiently. In this section we focus on a set of themes that are specific to the curation and reuse of *historical* data to support current and future science. While these themes echo some in the literature on the curation of contemporary scientific data, there are nuanced distinctions specific to retrospective data curation. Specifically, we will examine what our study revealed about (1) the perceived value of historical data for current and future scientific research and related projects; (2) challenges particular to recovering data from historical sources; and (3) ethical quandaries that arise in historical data recovery and reuse.

4.1 Data Value

The immediately evident value of historical data to contemporary science is supporting longitudinal analysis, such as complex modeling of natural systems over time. Our findings confirm this value, but also shed light on the nuances of the value of recovered data for different kinds of reuse: not only for longitudinal reuse within a domain, but longitudinal reuse across domains, for serving newly enabled research questions or methodologies, and for addressing social or infrastructural problems in public domains outside of academia. In summary, the cues to the value of historical data that we have identified so far stem from a wide variety of kinds of reuse by diverse communities:

- Reuse by researchers in the same domain, to support the study of novel research questions in light of new methodological opportunities or the advancement of contextual scientific understanding;
- Reuse in new, tangentially-related disciplines to study novel questions;
- Reuse to support meta-analysis, such as the historical study of science, or the evaluation of metrics or standards;
- Reuse by professional practitioners for decision-making, to inform policy and practice, and improve infrastructure or social conditions;
- Reuse by public communities to guide local decision-making or action.

We describe each of these varieties of reuse, which indicate different facets of the value of historical data, below.

Participants described needing long-term, observational data to make conclusions about scientific phenomena that require sweeping evidence, such as changes in biodiversity or climate conditions. For example, P17 describes the enduring value of observational data in ocean science:

> one of my main interests in data rescue and the reason that I do this work is that as an oceanographer, I was always data limited. Always. Right. There is not enough money in the world to get you out to sea often enough, in enough places, for long enough to get all the observational data that you want to have in order to describe an ecosystem, or even a place.

Informed by these data limitations, P17 wants to make historical ocean data available for researchers across multiple disciplines to enable longitudinal analysis and interdisciplinary modeling: “And in order to make an assessment of change in the ocean, you need time series data, and whether that’s biology or chemistry or meteorology, or what. And as you know, you can’t go back and re-collect the data. I mean, that’s why we do data rescue.”

Some participants described reusing archival data to address questions or test hypotheses newly enabled by technological developments or the progression of scientific knowledge and theory in certain domains. For example, P07, an assistant professor, described a recovery initiative “to understand how fish populations have changed in [one U.S. state] over time.” P07’s team is recovering historical fish survey records generated by a state agency to ask novel questions, such as, “can we use this data to accurately model the conditions that happened and therefore accurately model how things might happen in the future?”, and to test other ecological hypotheses related to species biodiversity and climate change. These are questions which were impossible to broach with these data in their original incarnation. Original lake inventories were done to support fishery management and to assess the success of fish stocking programs. The new questions being asked of these data are enabled by methodological advancement—particularly the development of computational and modeling techniques—but also by the advance of scientific theory in relevant ecological domains.

The same participant, speaking about a totally separate recovery initiative, described two further uses of the data: both to support pragmatic planning in the home collecting institution, and to support meta-analysis of scientific practice and standards. P07, speaking about the reuse of historical, paleontological data, describes how museum professionals plan to study “backlog fossils that are sitting in bags that have not been prepared…And they need to plan for a new building, and they want to use this data to better estimate the size of things for the new building.” This participant also has meta-analytic questions about the history of science and standardization, addressed through the paleontological data: “I’m interested in standards and how those change over time. I want to look at how the measurement system has changed over time”.

Our interviews show how recovered data are not only enabling new scientific inquiry but are being leveraged to address problems with immediate impact on professional practice, infrastructure, or public communities. One participant described their recovery and rescue work with agricultural data. P06, a data curator, discussed how “the history of the development of the crop varieties that we now use, is actually pretty valuable, especially as [person] says, under climate change, because …We’re going to have to develop new varieties to handle these conditions.” In this case, the recovered agricultural data are valuable because they can facilitate innovative crop production in the face of climate change.

In another case, a participant described pivoting the use of historical transportation data away from the original research motivation related to transportation *efficiency*, and toward consideration of transportation *equity*, reframing the data toward a newly perceived ethical imperative. P01, who is a data curator, detailed their experience working with transportation data, focusing on how data collected for one purpose can support another goal: “people weren’t thinking about it as equity. Even though it was a data set about transit in lower income neighborhoods in the United States, they weren't thinking about equitable access to transportation. We want to bring that out." Speaking further about this data, P01 discusses their intentions for the future use of this transportation data: "I want that work to have a positive impact on peoples' lives on peoples' health and decisions that get made about transportation in the future. It's much less about just saving the data to save the data. It's saving the data so it can be used to make people's lives better."

Other participants discussed not only reframing data toward novel, socially centered uses, but also reframing recovery initiatives themselves to center community needs that emerged during the initiative. P10 discussed their engagement with museums, associated Native American Graves Protection and Repatriation Act (NAGPRA) contacts, and plant legacy data, specifically describing the guidelines they assisted in creating, which outline access to collections based on Tribal protocols:

> But I worked with a NAGPRA person and the [museum] to determine the best home for some of the plant specimens, the photographs, because they're representative of multiple Tribal communities in the region, and things like that. But we also are working with them to open up access to the papers and then restrict it again and make it so that you need to talk to Tribal review boards to have access to it, because this would be considered sacred knowledge, where some of these plant locations are and things like that.

These procedures put Indigenous nations in control of who can access certain knowledge as based on Tribal sovereignty.

Another participant described refocusing a data rescue initiative to support the emergent, pragmatic needs of the data's originary community, in this case an Indigenous nation. P12, a librarian at an academic institution, described how a collaboration that started to address issues around archival descriptions, specifically the use of Native place names, became refocused on recovering data related to water usage in order to support an Indigenous nation's water claims. When it became clear to P12 that "some of the concerns that I thought that we would talk about, like Native place names, …that was not really a huge deal", P12 and their collaborators refocused their work on providing platforms to support this community: "They want the history, but they're also involved in legislation trying to get their water back. Making these materials discoverable, findable, having those conversations really focused us on the data related to the water history documents." P12's perspectives indicate a potential for pivoting data recovery work to be responsive to originary community needs. In this case, the effort was refocused on supporting Indigenous sovereignty at the intersection of data recovery and reuse. As discussed earlier, data recovery efforts can provide the long-term data needed to make broad conclusions about scientific phenomena, but they can also surface legacy data which may be valuable to vulnerable peoples and places, and their collaborators, furthering efforts that support Indigenous sovereignty and water claims.

Finally, data may be reused outside of research science to guide public and community action. P11, an associate professor, detailed how crowdsourced data collected to support research can be reused by people interested in maintaining their yards around the needs of local bird populations: "…the casual uses that people make – they use zebra [finch] data for things like deciding when to trim their trees, so they won't disturb breeding birds, right? Like very small scale homeownery, kind of landowner uses, those are also legit." People who are not scientists or researchers can also make use of scientific data as they care for their gardens.

4.2 Challenges Associated with Historical Data Recovery

It is well documented that scientists are reluctant to undertake data sharing and deposit even for their current data collections due to the difficulty and labor-intensiveness of data remediation, and the fact that the incentives at play in systems of scientific evaluation and credit make the publication of findings significantly more rewarding than the publication of data [29]. Historical data and data with differing original creators only compound this challenge. Our participants described prioritizing the curation and sharing of recent or current data, leaving no resources for the curation of older data. Given the priority that scientific research, funding, publishing, and evaluation place on novelty and innovation, there are also few incentives in the sciences for investigators to publish data that are not novel, or which originated from a different scientific author or investigator. Many participants working in positions and institutions dedicated to curation (rather than original scientific research) described a parallel conundrum. They expressed that it is difficult to justify digitization work to institutional leaders and funding agencies as work on legacy data, even when it is known to have high research value, is seen as in competition with work on more current data, which is considered burdensome enough on its own. Justification often comes from the data being desired for research, but the invisibility of legacy data leads to a perception that it is irrelevant. P20, a metadata manager, described this problem: "sometimes it's a matter of does, is someone wanting this data now? And then I can go to the powers that be and say, someone wants this data, can we have resources for it? But if nobody, it's a catch 22: If no one knows it's there, no one's going to ask for it."

Even when the funding and time are available for the recovery of historical data, the data itself is often difficult to extract: legacy data is saved in file types or formats that no longer exist, or are inaccessible to modern tools used for data curation and management. P01 described a specific scenario wherein data had to be manually scraped from webpages:

> …they've asked us to go back and preserve the legacy data and help pull it off the webpage, because all these files are attached to HTML pages. And some of them aren't actually saved anywhere in a drive that we can find. So, the only file available is an attachment to an HTML page. And we all know that that's a nightmare waiting to happen. So, we're often scraping hand, scraping links from web pages and downloading stuff.

These factors together mean that recovery of data from historical sources is often time- and labor-intensive and is often an additional or marginal side project for curators rather than the central focus of any dedicated position. Necessarily, shortcuts are sometimes taken, resulting in suboptimal or "good-enough" recovery efforts. P01 states:

> And oftentimes, what we would like to make that into a time series or something else, we don't have the time for that so it's: save each survey as we can, pull up documentation off of the web page, and then preserve it and make it public, with the caveat that the public presents us with a request and a good use case, we will do more work. But oftentimes, it's how do we make the most efficient use of our time to make this public in at least an open format.

4.3 Ethics of Data Recovery

Alongside larger challenges of data reuse, participants also observed ethical concerns related to data recovery. Ethical questions emerged in two strands. The first considered questions around how data condenses from a larger relationship to historical inequities in naming choices and in research practice. The second category of ethical concern reflected questions of researcher knowledge production and the intent behind data reutilization. Participants expressed concerns about using data collected from vulnerable or marginalized populations, particularly where those populations do not justly or equitably benefit from the outcomes of the research, or where the same communities do not have control over how data are ultimately used. P06 stated:

> …it's also a priority for [organization], because it's a priority for this president, to make sure that we're doing things equitably. And so, to my mind, that means not just — are we serving those populations with our programs fairly? But are we making sure that if we're collecting data from those populations, on their farming practices, or on their use of nutritional benefits, for example, is the data being used fairly? And do they have some say in how their data is going to be used?

For the latter concern, P05 tied the impact of data recovery to a larger exploration of artificial intelligence (AI) research and ethics. Expressing concern that scientists working with AI techniques are frequently guilty of "introducing bias into algorithms," P05 suggested that the availability of data for recovery and reuse should not equate to free, unregulated use. Highlighting the value of information professionals and librarians in ethical AI, P05 asserted that ethical recovery and reuse of data should include funded "training" and methods for identifying "expertise" within data curation work. In notable divergence, P11 argued that this type of credentialing raises its own ethical concerns given what they see as chronic acts of dismissing community-based science and the use of data by non-experts as irrelevant to the advancement of knowledge. Overall, these ethical quandaries probe issues surrounding who benefits from the use of recovered data and how academia understands recovered data as a method of knowledge production.

5 Discussion and Future Work

This research illuminates new aspects of the value of historical data recovery and reuse. We have offered a very preliminary framework of kinds of reuse, organized around diverse reuse communities—including domain scientists and researchers in new disciplines to support the study of new research questions in light of new methodological opportunities or the advancement of scientific understanding, or for cross-disciplinary analysis; by researchers in different disciplines doing meta-analysis, including social and historical studies of science; by professional practitioners, working in professional domains or knowledge institutions; and by public communities, to guide decision-making or local action. There are valuable prior frameworks of scientific data reuse [17], and separate frameworks related to the value and impact of archival data [30]. Prior work on data reuse has mainly considered contemporary scientific data (rather than historical data). Prior work on archival impact has largely omitted the scientific applications of archival or cultural data. Because we developed our preliminary framework through inductive coding, we have not yet aligned our findings with prior frameworks, but we believe our findings will round out prior frameworks with an emphasis on historical data and data across disciplines.

Historical data recovery poses myriad challenges. Many of them echo factors from the extensive prior literature on scientific data curation, data sharing, and data reuse. For example, as has been widely documented in prior work on scientists' data practices [31], many researchers who recover and recreate historical datasets are reluctant to openly share data after its recovery, due to concerns about how others will perceive the data's quality, the additional labor of preparing data for sharing (e.g., of providing adequate documentation to support independent understandability of the data), or the necessity of retaining competitive research advantages. Even the familiar challenges, however, are compounded by the fact that data stemming from recovery initiatives are divorced from their original creators and contexts. They may never have been shared originally *as data*, since these datasets have often been manually reconstructed from analog or digital sources in fundamentally different formats: from narrative text of field notes, from the coded fields of ships logs, from the captions or labels of images or graphs, or from tabular data in different units of measurement. Because they are often being repurposed and recontextualized, these data require a certain level of expertise to be constructed in the first place without introducing errors of historical misinterpretation. In fact, these data may never have been purposefully shared *at all*, having been recovered and disseminated after a scientist's retirement or decades after the work was done, without the data creators' knowledge or consent.

In addition, many historical datasets were collected under paradigms of scientific observation and data collection that do not meet contemporary ethical standards of research—in terms of how they exploit historical or current communities and their resources, or how they represent or identify entities within the data in offensive or outmoded ways. For this reason, data being recovered and reused to support new scientific research have much in common with data leveraged in humanistic, historical, and anthropological research, derived directly from historical primary sources and gathered from archival collections built through exploitative or colonial collecting practices. They also share characteristics with qualitative and social scientific data, which are notoriously fraught with potential risks to the privacy, confidentiality, and wellbeing of human-subjects research participants. These branches of data curation research—across the sciences, social sciences, and humanities—rarely intersect. Future work aims to identify opportunities for more well-established practices and discourse around data reuse across the humanities and social sciences to inform scientific data recovery. There is also a need to bring research on archival data recovery into conversation with the theory and practice related to collections-as-data [32, 33] and ethical implications in archives, and library and information science. Mapping our findings to extant frameworks of data curation activities [34] as part of this future work will also help identify gaps in current data curation training and practice relative to retrospective curation.

Finally, future work on this research will aim to produce and disseminate guidance for archivists, librarians, and data curators who work with and preserve historical materials, to support the extraction and reuse of useful scientific data as part of broader digital curation processes, or to support individual scientists' efforts. This work is situated in a broader need to explore the distinctions between how research scientists go about data recovery, to find answers to specific questions, and how professional curators approach data recovery to support open-ended possibilities of reuse. Our future work, looking at data curation across a broader spectrum of disciplines, aims to shed light on this question and the convergence of curatorial roles.

References

1. Shiue, H.S.Y., Clarke, C.T., Shaw, M., Hoffman, K.M., Fenlon, K.: Assessing legacy collections for scientific data rescue. In: Toeppe, K., Yan, H., Chu, S.K.W. (eds.) iConference 2021. LNCS, vol. 12646, pp. 308–318. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71305-8_25
2. Choudhury, S.: Data at risk and research libraries. In: AGU Fall Meeting Abstracts, IN21E-01 (2017)
3. Mayernik, M.S., et al.: Stronger together: the case for cross-sector collaboration in identifying and preserving at-risk data. Figshare, 1 (2017). https://doi.org/10.6084/m9.figshare.4816474.v1
4. Downs, R.R., Chen, R.S.: Curation of scientific data at risk of loss: data rescue and dissemination. Columbia University Academic Commons (2017). https://doi.org/10.7916/D8W09BMQ
5. Cragin, M.H., Heidorn, P.B., Palmer, C.L., Smith, L.C.: An educational program on data curation. Poster Presentation. American Library Association, Washington, D.C., 25 June 2007. https://hdl.handle.net/2142/3493
6. Higgins, S.: The DCC curation lifecycle model. Int. J. Digit. Curation **3**(1), 134–140 (2008). https://doi.org/10.2218/ijdc.v3i1.48
7. Vearncombe, J., Riganti, A., Isles, D., Bright, S.: Data upcycling. Ore Geol. Rev. **89**, 887–893 (2017). https://doi.org/10.1016/j.oregeorev.2017.07.009
8. Mayernik, M.S., Breseman, K., Downs, R.R., Duerr, R., Garretson, A., Hou, C.-Y.: Risk assessment for scientific data. Data Sci. J. **19** (2020). https://doi.org/10.5334/dsj-2020-010
9. Janz, M.M.: Maintaining access to public data: lessons from data refuge, 5 March 2018. https://doi.org/10.31229/osf.io/yavzh
10. McGovern, N.Y.: Data Rescue. ACM SIGCAS Comput. Soc. **47**(2), 19–26 (2017). https://doi.org/10.1145/3112644.3112648
11. Allen, L., Stewart, C., Wright, S.: Strategic open data preservation: roles and opportunities for broader engagement by librarians and the public. Coll. Res. Libr. News **78**(9) (2017). https://doi.org/10.5860/crln.78.9.482

12. Walker, D., Nost, E., Lemelin, A., Lave, R., Dillon, L.: Practicing environmental data justice: from DataRescue to data together. Geo Geogr. Environ. **5**(2) (2018). https://doi.org/10.1002/geo2.61
13. Brunet, M., Jones, P.: Data Rescue Initiatives: bringing historical climate data into the 21st century. Climate Res. **47**(1), 29–40 (2011). https://doi.org/10.3354/cr00960
14. Wyborn, L., Hsu, L., Lehnert, K., Parsons, M.A.: Guest editorial: special issue rescuing legacy data for future science. GeoResJ **6**, 106–107 (2015). https://doi.org/10.1016/j.grj.2015.02.017
15. Fallas, K.M., MacNaughton, R.B., Sommers, M.J.: Maximizing the value of historical bedrock field observations: an example from Northwest Canada. GeoResJ **6**, 30–43 (2015). https://doi.org/10.1016/j.grj.2015.01.004
16. Specht, A., Bolton, M., Kingsford, B., Specht, R., Belbin, L.: A story of data won, data lost and data re-found: the realities of ecological data preservation. Biodivers. Data J. **6** (2018). https://doi.org/10.3897/bdj.6.e28073
17. Pasquetto, I.V., Randles, B.M., Borgman, C.L.: On the reuse of scientific data. Data Sci. J. **16**(8), 1–9 (2017). https://doi.org/10.5334/dsj-2017-008
18. Borgman, C.L.: Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press, Cambridge (2015)
19. Palmer, C.L., Weber, N.M., Cragin, M.H.: The analytic potential of scientific data: understanding re-use value. Proc. Am. Soc. Inf. Sci. Technol. **48**(1), 10 (2011). https://doi.org/10.1002/meet.2011.14504801174
20. Meystre, S.M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., Lehmann, C.U.: Clinical data reuse or secondary use: current status and potential future progress. Yearb. Med. Inform. **26**(01), 38–52 (2017)
21. Wolinetz, C.D., Collins, F.S.: Recognition of research participants' need for autonomy: remembering the legacy of Henrietta Lacks. JAMA **324**(11), 1027–1028 (2020)
22. Marsh, D.E., Punzalan, R.L., Johnston, J.A.: Preserving anthropology's digital record: CoPAR in the age of electronic fieldnotes, data curation, and community sovereignty. Am. Arch. **82**(2), 268–302 (2019). https://doi.org/10.17723/aarc-82-02-01
23. Mannheimer, S.: Data curation implications of qualitative data reuse and big social research. J. eSci. Librariansh. **10**(4), 5 (2021). https://doi.org/10.7191/jeslib.2021.1218
24. Duke, C.S., Porter, J.H.: The ethics of data sharing and reuse in biology. Bioscience **63**(6), 483–489 (2013)
25. Voytek, B.: The virtuous cycle of a data ecosystem. PLoS Comput. Biol. **12**(8), e1005037 (2016)
26. Curty, R.G.: Beyond "data thrifting": an investigation of factors influencing research data reuse in the social sciences. Doctoral dissertation, Syracuse University (2015)
27. Guba, E.G., Lincoln, Y.S.: Competing paradigms in qualitative research. In: Handbook of Qualitative Research, vol. 2, no. 163–194, p. 105 (1994)
28. Boeije, H.: A purposeful approach to the constant comparative method in the analysis of qualitative interviews. Qual. Quant. **36**(4), 391–409 (2002)
29. Acord, S.K., Harley, D.: Credit, time, and personality: the human challenges to sharing scholarly work using Web 2.0. New Media Soc. **15**(3), 379–397 (2013)
30. Marsh, D.E., Punzalan, R.L.: Studying and mobilizing the impacts of anthropological data in archives. In: Crowder, J.W., Fortun, M., Besara, R., Poirier, L. (eds.) Anthropological Data in the Digital Age, pp. 163–183. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-24925-0_8
31. Fecher, B., Friesike, S., Hebing, M.: What drives academic data sharing. PLoS ONE **10**(2), e0118053 (2015). https://doi.org/10.1371/journal.pone.0118053

32. Padilla, T., Allen, L., Frost, H., Potvin, S., Roke, E.R., Varner, S.: Always already computational: collections as data: final report. University of Nebraska Digital Commons. University of Nebraska, Lincoln (2019). https://digitalcommons.unl.edu/scholcom/181/. Accessed 18 Sept 2022
33. Coleman, C.N.: Managing bias when library collections become data. Int. J. Librariansh. **5**(1) (2020). https://doi.org/10.23974/ijol.2020.vol5.1.162
34. Lafia, S., Thomer, A., Bleckley D., Akmon, D., Hemphill, L.: Leveraging machine learning to detect data curation activities. In: 2021 IEEE 17th International Conference on eScience (eScience), pp. 149–158 (2021). https://doi.org/10.1109/eScience51609.2021.00025