Context from the data reuser's point of view

By Ixchel M. Faniel, Rebecca D. Frank, Elizabeth Yakel

Abstract

Purpose – Taking the researchers' perspective, this paper examined the types of context information needed to preserve data's meaning in ways that support data reuse.

Design/methodology/approach – This paper is based on a qualitative study of 105 researchers from three disciplinary communities: quantitative social science, archaeology, and zoology. The study focused on researchers' most recent data reuse experience, particularly what they needed when deciding whether to reuse data.

Findings – Findings show researchers mentioned twelve types of context information across three broad categories: 1) data production information (data collection, specimen and artifact, data producer, data analysis, missing data, research objectives), 2) repository information (provenance, reputation and history, curation and digitization), and 3) data reuse information (prior reuse, advice on reuse, terms of use).

Originality/Value – This paper extends digital curation conversations to include the preservation of context as well as content to facilitate data reuse. When compared to prior research, findings show that there is some generalizability with respect to the types of context needed across different disciplines and data sharing and reuse environments. It also introduces several new context types. Relying on the perspective of researchers' offers a more nuanced view that shows the importance of the different context types for each discipline and the ways disciplinary members thought about them. Both data producers and curators can benefit from knowing what to capture and manage during data collection and deposit into a repository.

Keywords Data curation, Data sharing, Data reuse
Paper type Research paper

Suggested Citation:

Faniel, I., Frank, R. and Yakel, E. (2019), "Context from the data reuser's point of view," *Journal of Documentation*, Vol. 75 No. 6, pp. 1274-1297. https://doi.org/10.1108/JD-08-2018-0133

1. Introduction

Context is a critical component for data reuse. Defined as “the interrelated conditions in which something exists or occurs” (Merriam Webster Online Dictionary, http://www.merriam-webster.com/dictionary/context), the technical aspects of context necessary for reliable, long-term access to digital content (e.g. cultural objects, data, images, software, etc.) have been explored in the digital preservation and curation literature. Less attention has focused on the context necessary for preserving content’s meaning to support reuse. Yet, preserving meaning has become increasingly important as scholarly communities look to share and reuse research data to replicate and reproduce research as well as perform novel inquiries. Although researchers have proposed frameworks (Baker and Yarmey, 2009; Beaudoin, 2012a, 2012b; Lee, 2011), engaging with data reusers to specify the context necessary to support meaning making has occurred less frequently.

Lee’s (2011) framework for context information in digital collections outlines three context types for any given target entity (e.g., object, data, record): 1) representations or relational entities of the target entity, 2) factors external to but potentially acting upon the target entity, 3) perceptions of the user of a target entity. Drawing from the cataloging and classification, archives, and information needs literatures, he presents a framework comprised of nine classes of contextual entities. Although Lee provides a broad view of context, the framework primarily focuses on the context information necessary to ensure that digital objects can be rendered over time. According to Lee (2011, p. 104), the entities provide an exhaustive documentation of a digital object and the framework “is intended to inform the creation, capture and curation of the contextual information within a repository, which can help to understand, make sense of, analyze and use a particular target digital object.” Albeit a valid approach and important contribution, Lee’s framework primarily speaks to digital curators’ activities, providing very few details about the context necessary for meaning making during reuse.

Based on a review of the digital preservation literature, Beaudoin (2012a) develops a multi-dimensional framework. The framework goes beyond technical aspects of context to include utilization, physical, intangible, curatorial, authentication, authorization, and intellectual aspects. Taken together these elements enable “successful retrieval, assessment, management, access, and use of preserved digital content” (Beaudoin, 2012a, Abstract). In subsequent work, Beaudoin (2012b) generates a questionnaire intended to generate a set of metadata elements to describe an object’s context according the framework. Unfortunately, the questions fall short of defining audience needs beyond several high-level categories, such as educational, leisure, legal, medical, and youth.

Expressing user and reuser needs at this level of abstraction is common. Studies acknowledge that capturing data context is difficult, noting context’s tacitness, the human labor costs required to create and manage it, and data creators’ and information professionals’ lack of expertise and knowledge about what should be captured (Berg and Goorman, 1999; Birnholtz and Bietz, 2003; Edwards et al., 2011). Interestingly, few studies in research and practice have asked researchers about their data reuse needs in general or their needs for context information about the data specifically. Yet, a key premise of the Open Archival Information System (OAIS) reference model for the curation and preservation of digital objects is identifying designated communities to determine whether the context information needed to support reuse is being provided (Consultative Committee for Space Data Systems, 2012). Many librarians and archivists take issue with the term ‘designated communities’, which can encompass a broad public making the task of preserving meaning such that it can be understood by everyone unfathomable (Bettivia, 2016). Even for those who have a more focused designated community in mind, it is rare that the community is defined explicitly, and its members’ input systematically collected and incorporated into digital repository design (Bettivia, 2016). For these reasons, we draw from the data sharing and reuse literature, particularly Chin and Lansing’s (2004) study, which offers the most comprehensive view of different context types a data reuser may require.

Using participatory analysis and design sessions, Chin and Lansing (2004) asked biologists to describe different scenarios of collaborative data sharing and reuse, including the kinds of information and computational capabilities needed to support data practices within the Biological Science Collaboratory (BSC). Drawing from the context types Chin and Lansing identified, we consider whether their model can be generalized to researchers in three other disciplines, whose data reuse is not the result of joint work in a collaboratory – quantitative social science (hereafter referred to as social science), archaeology, zoology. The research questions posed are:

1. What types of context information do researchers need when deciding whether to reuse data?
2. How do researchers’ need for different types of context information vary across disciplinary communities?

Our findings confirm and expand Chin and Lansing’s conclusions about context, particularly as it relates to what researchers need to evaluate data for reuse. Before discussing our findings in detail, we present a literature review drawing from Chin and Lansing and other data sharing and reuse studies followed by our research methodology. We then discuss our findings and the implications they have for data producers and repository staff capturing and curating context information for reuse.

2. Literature review

Focused on collaborative data sharing and reuse, Chin and Lansing identified two broad categories of context biologists working in the BSC needed to describe how data were created, interpreted, and applied: scientific and social. These categories were the scientific context, which supported biologists’ research with the data, and the social context, which facilitated biologists’ collaboration with each other. Given that our study examines data reuse practices that are not the result of joint work within a collaboratory, we draw from four key context types Chin and Lansing identified that focus on the scientific context: 1) general data set properties, 2) experimental properties, 3) data provenance, and 4) analysis and interpretation.

2.1 General data set properties

General data set properties include the data owner, data creation date and time, modification date and time, file name, file size, data format, and description. Although Chin and Lansing do not describe the role of these kinds of metadata beyond acknowledging that one size does not fit all, some studies have found that these properties can be significant for data reuse. The most widely referenced property in data reuse studies is the data owner and this most often refers to the person who produced the data. Data reusers often reference the data producer's name to recall knowledge about training, institutional affiliations, and publications (Zimmerman, 2008). That information is used to assess the data producer's research competence and reputation, which is often the first step in assessing the data's trustworthiness (Berg and Goorman, 1999; Van House, 2002; Van House et al., 1998). Data reusers rely on data creation dates to develop sampling frames for their studies (Zimmerman, 2007) as well as identify points of change in the data production context related to internal and external factors, which influence how data should be interpreted (Birnholtz and Bietz, 2003). Data formats can indicate the degree to which the data are accessible, given the software data reusers have available or are skilled at using (Yoon, 2016) and whether the data are available in physical and/or digital forms (Carlson and Anderson, 2007). Lastly, reusers read general descriptions about a dataset to help identify data for more detailed evaluation (Faniel and Jacobsen, 2010).

2.2 Experimental properties

Experimental properties describe the "conditions and properties of the scientific experiment that generated or is to be applied to the data" (Chin and Lansing, 2004, p. 410). For data reuse to occur, the research design needs to be recorded consistently and thoroughly. Earthquake engineering researchers reusing colleagues' data to validate numerical computational models required documentation that detailed experimental parameters (i.e. test set-ups) and procedures, so assessments could be made about whether data were relevant, understood, and trustworthy (Faniel and Jacobsen, 2010). Before using data from an electronic detector, astronomers needed to know the algorithm used to calibrate the device and before using survey data, social scientists needed to know details about survey design, administration, processing, and analysis (Carlson and Anderson, 2007). These details are important, because data collection processes are highly individualized across quantitative and qualitative disciplines (Carlson and Anderson, 2007). Even in a small nine-person materials science lab, there was variation in what scientists deemed important to record in a shared laboratory notebook, which resulted in the absence of key experimental details and prevented data reuse (Akmon et al., 2011). Trying to create consistent and thorough documentation of these processes is both time consuming and difficult. Darch et al. (2015, p. 68) found these differences were due in part to "different and constantly changing configurations of social, material, and scientific resources to help them accomplish different steps of the workflow."

2.3 Provenance

Discussions of data provenance extend the definitions found in the archival and management information systems literatures. The archival tradition defines provenance as source, which can either be a data creator or a data repository (Pearce-Moses, 2005). Fear and Donaldson (2012) use this definition when describing how proteomics researchers used provenance (source) information as a cue to data quality, to assess data producer reputation, and as an indicator of expertise. Similarly, to facilitate trust in data, scientists in the SkyProject record the provenance as source in terms of the instrumentation and its settings so they and their colleagues can later verify their own and their peer's experiments (Carlson and Anderson, 2007). In the management information systems literature, Wang and Strong (1996) describe provenance as source, but also as the ability to link information to a source or provenance as traceability. They then tie traceability to an assessment of data quality. Similarly, in terms of research data, Chin and Lansing (2004, p. 410) describe data provenance as the "relationship of data to previous versions and other data sources." They suggest that being able to trace data's historical lineage allows researchers to judge data's credibility and reliability. Buneman, Khanna, and Tan (2001) share this view of data provenance as the lineage or pedigree of data. For them, data provenance is "the description of the origins of a piece of data and the process by which it arrived in a database" (Buneman et al., 2001, p. 316). They argue that this is important for determining the accuracy and currency of data. These definitions align with other definitions of provenance as traceability (Huang, 2015; Wang and Strong, 1996).

2.4 Analysis and interpretation

Chin and Lansing (2004) show how the BSC facilitates collaborative analysis and interpretation. Working together, biologists developed findings and drew conclusions by sharing and building upon each other's understandings of the data. To support this effort the BSC went beyond capturing structured metadata "to capture descriptions, critiques, interpretations, experiences, and general knowledge…such as elaborate descriptions of a data set, instructions on how the data set may be applied, identification of errors or anomalies in the data, and results of analysis using the data" (Chin and Lansing, 2004, p. 413). Data reusers have consistently found that understanding how and why data are constructed in an original study is difficult and time consuming (Rolland and Lee, 2013). Confusion and disagreement about the processes (e.g. cleaning, manipulating, and analyzing data) can halt reuse (Yoon, 2016). Given that research studies are not without problems, it is not typically data deficiencies that hinder reuse, but the lack of documentation about the context in which the data were collected and analyzed that causes difficulties. For instance, earthquake engineering researchers referred to documentation to see what led to data deficiencies and what, if anything, was done to compensate for them, to increase their understanding of and trust in the data (Faniel and Jacobsen, 2010).

Drawing from these findings in the data sharing and reuse literature, the purpose of this study was to identify the dynamics of reuse practices across the social science, archaeological, and zoological communities. We were particularly interested in how researchers within these communities viewed context and under what circumstances it contributed to their reuse practices, specifically their decisions about whether to reuse data.

3. Research methodology

This article is based on 105 interviews and observations with researchers from three disciplinary communities: social science, archaeology, and zoology. The interviews and observations were done from 2011-2013 as part of a larger research project. A major objective of the project was to examine contextual factors affecting data reuse in these three disciplinary communities.

3.1 Data collection: Participants and partner organizations

Convenience and snowball sampling were employed via three recruitment methods: 1) working with our research partners to recruit a diverse group of interviewees, 2) recruiting at disciplinary conferences, and 3) asking participants to suggest other potential interviewees. Data collection continued until the team felt it had achieved data saturation. All interviews and observations lasted 30-60 minutes and participants were offered $25 for their participation. The Institutional Review Board at the third author's university reviewed and approved the study.

3.1.1 Quantitative social scientists: Inter-university Consortium for Political and Social Research. The Inter-university Consortium for Political and Social Research (ICPSR) was founded in 1962. ICPSR is a global leader in social science data stewardship and maintains a data archive of more than 500,000 files of social science research (The Inter-university Consortium for Political and Social Research, n.d.). Comprised of more than 700 academic institutions and research organizations, the consortium serves a diverse community of researchers. We spoke with 21 expert and 22 novice social scientists who sought statistical data on such diverse topics as political movements, criminal justice, and health care. Three of the novice social scientists interviewed were aspirational rather than actual data reusers, which meant they had not reused data, but wanted to in the future and were able to talk about data of interest. Of the 43 social scientists, 42% identified as faculty, 14% as research scientists, 47% as students, and 5% as future/prospective graduate students. Of the 43 participants, 58% reported collecting their own data. Thirty percent reported reusing data obtained directly from colleagues, while 93% used online repositories and/or websites.

3.1.2 Archaeologists: Open Context. Open Context is an open access data publication platform, maintained by the Alexandria Archive Institute ("About Open Context," 2015). Founded in 2007, Open Context seeks to improve standards that support data reuse and provide a platform for archaeologists to share and reuse a wide variety of data (Faniel et al., 2013; Kansa et al., 2007). The repository supports data sharing as a new form of publication and a complement to conventional forms ("About Open Context," 2015). Of the 22 archaeologists interviewed, 14% were curators, 41% faculty, 18% faculty-curators, 9% research scientists, and 18% graduate students. Of the 22 interviewed, there was one aspirational and 21 actual data reusers and 73% reported collecting their own data. Depending on their research interests, archaeologists relied on particular aspects of material culture, such as physical objects, textual descriptions, images of objects, and data about the originating site as well as faunal remains and osteological data. Sixty-four percent reported reusing data obtained from colleagues, 41% from online repositories and/or websites, 77% from museums and archives, and 50% from publications.

3.1.3 Zoologists: University of Michigan Museum of Zoology. Founded in 1885, the University of Michigan Museum of Zoology (UMMZ) collections include about 15 million specimens, including mammals, birds, amphibians and reptiles, fishes, mollusks, mites, and insects (University of Michigan Museum of Zoology, 2015). UMMZ supports scientists' and students' research across a number of fields, and contributes data to several discipline-specific digital repositories, such as VertNet and the Global Biodiversity Information Facility (GBIF). These various access points provide researchers access to UMMZ data. UMMZ also maintains its own specimen databases and holds a substantial number of field notebooks documenting the collection of its specimens. We conducted 27 interviews and observed 13 zoologists. Of the 40 zoologists, 8% identified as curators, 23% as faculty, 13% as faculty-curators, 15% as postdoctoral researchers, 3% as research scientists, and 40% as graduate students. All 40 zoologists were actual data reusers and 63% reported collecting their own data. Twenty percent reported reusing data obtained directly from colleagues, 90% from online repositories and/or websites, 95% from museums and archives, and 28% from publications.

All three disciplinary communities used both qualitative (in its broadest sense) and quantitative data in their work. In each discipline, particularly quantitative social science, there were individuals whose work was exclusively quantitative; others used qualitative data to generate quantitative data, as was the case for some of the archaeologists and zoologists. None of the participants were qualitative researchers in the sense that they employed interpretivist approaches to data. This is a limitation of our study.

3.2 Data Analysis

Audio recordings from the interviews and observations were transcribed, then coded using the qualitative data analysis package NVivo. Combining deductive and inductive approaches, research team members began with a code set developed based on the interview protocol and data sharing and reuse concepts from the literature (Miles and Huberman, 1994). Two team members initially coded each wave of data collection, to establish inter-rater reliability. Scott’s pi, a method for calculating agreement among coders (Scott, 1955), was used. The results for each wave were: expert social scientists (0.77), novice social scientists (0.88), archaeologists (0.73), zoologists’ interviews (0.74), zoologists’ observations (0.88).

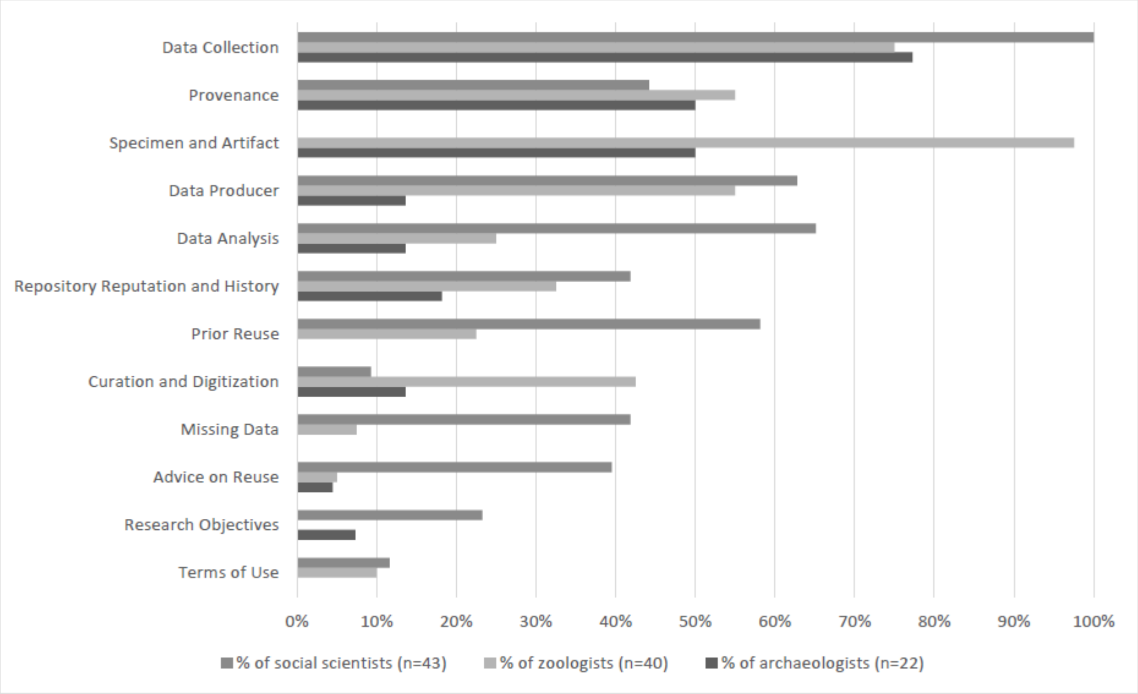
This article is based on a subset of data. We queried the interview data based on codes and sub-codes in the following areas: context, data reuse, dissemination of data, interaction of stakeholders, and trust. Next, we developed a second code set focusing on three primary areas: 1) the context the reuser needed, 2) the reasons the reuser needed context, and 3) the places the reuser went to get context. These codes were developed based on concepts from the data sharing and reuse literature as well as themes identified during the first round of data analysis. We then divided the subset of data selected for this second cycle of analysis among the three authors and coded using an iterative process that included regular meetings to discuss coding and identify emergent themes not present in the code set.

Table 1. Types of context information, definitions, and examples

Context Types	**Definitions**	**Examples**
Data Collection	Data acquisition processes, techniques, or issues.	Study design, site selection, location, sampling, measurement, data management, file format, tools and instrumentation
Specimen and Artifact	Specimen: Zoological collection and/or curation information on animals captured or recovered in the field and/or genus/species information. Artifact: Animal, human, or material remains uncovered during an archaeological field survey or excavation.	Specimen species/genus name; location, physical characteristics (e.g., size, condition); identification information (e.g., accession number or identifier) Artifact identification, time period, location (strata), relationship to other artifacts (stratigraphy)
Data Producer	The individual or organization that created or compiled the data.	Specific name or characteristics of the person or institution involved in data creation/compilation
Data Analysis	Data preparation or interpretation processes, techniques, or issues leading to conclusions.	Cleaning, transforming, aggregating, calculating, or coding data; creating DNA sequence or radiocarbon dating data for specimens or artifacts.
Missing Data	Instances when there were no data stored for variables or observations.	Lists of variables or observations with no data, reasons why there are no data, quantification of how much data are not present, how to deal with lack of data
Research Objectives	Goals or research questions associated with a research study.	Research questions, hypotheses, or models
Provenance	Source of the material or traceability/action taken on the data overtime.	Creator, contributor, repository name (source) or data analysis or cleaning information, curation processes (traceability)
Repository Reputation and History	Generally held beliefs and knowledge of past events about an institution holding data, specimens, or artifacts.	Names of specific physical or online repositories, comments on parent institutions or affiliations, staff (size, expertise); reputation;

		collection scope; or curatorial processes
Curation and Digitization	Descriptions of how data, specimens, or artifacts were curated or digitized, including the people, functions, and/or services associated with these activities	Curation: Data management or preservation processes of physical materials or digital information (e.g. paper, specimens) and other value added curation practices, such as description, documentation, verification, access provision or preservation Digitization: how, why, when, and what was digitized; standards used; digitization quality
Prior Reuse	Persons engaged in or publications (e.g., data journal or commentaries) recounting prior reuse of data	Citations to data reuse studies, data critiques, etc.; References to others reusing data
Advice on Reuse	Guidance provided to support data reuse	Recommendations or tips on aspects of data reuse from published studies, peers, mentors, or repository staff
Terms of Use	Data producer or repository rules or regulations surrounding data reuse	Copyright; access restrictions; special certifications for reuse; privacy or confidentiality of human subjects' identity or location of artifacts and specimens; anonymization rules

Figure 1. Context types by discipline



Context types study participants mentioned needing or wanting when deciding to reuse data. Based on interviews and observations with social scientists (n=43), zoologists (n=40), and archaeologists (n=22).

4. Findings

Findings showed researchers mentioned twelve types of context information that we organized into three broad categories (Table 1): 1) data production information (data collection, specimen and artifact, data producer, data analysis, missing data, research objectives), 2) repository information (provenance, reputation and history, curation and digitization), and 3) data reuse information (prior reuse, terms of use, advice on reuse). Although many of these context types have been mentioned in the literature, our findings present a more nuanced view from the perspective of researchers in three disciplines, which demonstrate different degrees of reliance on each context type (Figure 1). Findings showed not only the importance of the different types of context for each discipline, but also the ways disciplinary members thought about the different types. The researchers' disciplinary needs are shown through illustrative examples and the similarities and differences across disciplines are highlighted.

4.1 Data production information

Data producers are the primary source of data production information. Four of the six types of data production information (data collection, specimen and artifact, data producer, and data analysis) were among the most highly sought-after context types across all three disciplines. Majorities of researchers in all disciplines mentioned data collection information. The lower percentage of zoologists and archaeologists discussing data collection information was likely due to more focused comments on data collection types specific to their disciplines: specimens and artifacts. Specimen and artifact information was the third most frequently mentioned context type. Data producer information was fourth. Data analysis information followed with social scientists mentioning it most frequently. Missing data and research objectives were mentioned less frequently across all disciplines.

4.1.1 Data Collection. Researchers across the disciplines wanted information about data collection procedures. All social scientists, 77% of archaeologists, and 75% of zoologists mentioned data collection information. Social scientists reusing survey data wanted to know how the questionnaires were designed and administered to determine data's relevance. Interested in "serious crimes", Social scientist 33 explained that how survey questions were asked signaled a study's definition and measurement of criminal activity.

> So a lot of datasets will say that they have some crime information, and you look into it, and it's… "Have you gotten into a fight? Did you skip school?"... we're looking for something a little bit more serious than that. "Have you broken into a building with the intention of stealing something? Have you injured someone in a fight seriously enough that they had to seek medical treatment or go to a hospital?" (Social scientist 33).

Social scientists also sought data collection information about the population under study and the sampling procedures. Limiting his study of serious crimes to the Asian community, Social scientist 33 was concerned about whether he could find data with a sample size large enough to yield statistically significant results.

Data collection information also reflected a data producer's research interests and oftentimes researchers used it to assess whether the data would meet their reuse needs. Archaeologist 07 described differences in data collection procedures during an excavation as "digging with a tractor vs. a dental pick". Similarly, Archaeologist 01 explained that differences among excavation procedures resulted in different data granularities and reusers needed to know the procedures to assess fitness for reuse.

> So, they're walking huge tracts of land, but they're only hitting big things. They'll hit a site, but they'll walk by little tiny sherd scattered things. So you kind of need to know that. I've heard of things like shoulder surveys, where they literally walk side by side and pick those little things, but then, again, you've only, you're doing a very narrow tract. So there are procedures (Archaeologist 01).

For reusers interested in combining data from different studies, data collection information was used to determine feasibility. Zoologist 11 used the times of the year and the day, and weather conditions from historical bird survey data to determine whether they matched current standards or should be removed from analysis.

> In the bird survey world, if the survey was one of the extremes of the survey season that is not currently within the season that we normally do surveys now, or if the time of day was really weird, or if the weather recorded was really bad... If essentially, they were doing the surveys under conditions or in different ways that did not match our current standards, and if it makes sense to essentially filter that data out without biasing what you're filtering (Zoologist 11).

Social scientists shared similar sentiments, but relied on the questions asked and how responses were measured to determine whether they could combine data from multiple studies. Archaeologists also combined data across studies, but often they were hampered by lack of consistency. Archaeologist 02, who wanted to analyze data at as high a resolution as possible, discussed the challenges related to imprecision in archaeologists' data collection descriptions.

> …when you identify an animal bone, you can either identify the species, you can say it's a dog, a canis familiaris, or you can just say that it's a canine. It could be a fox, a wolf, or a dog…So, rectifying that data is time consuming and complex and you have tables go through on a case by case basis, and there's no real parsing routines that you can sort of develop that can go through that for you…it's a real technical problem and quite often it results in dismissing data that's not the real, the taxonomical resolution that you want, and so you're always sort of trading off between...high resolution and sort of coarse resolution when you're combining data (Archaeologist 02).

4.1.2 Specimen and artifact information. Almost all zoologists (98%) and half of the archaeologists (50%) discussed specimen and artifact information respectively. Zoologists mentioned needing taxonomic identification of specimens at the species level (i.e., species name) and the location of their capture (i.e., latitude and longitude coordinates) most frequently. Experienced zoologists used this information in decisions about the relevance and accuracy of data. Given his expertise, Zoologist 08 used location information to verify the accuracy of taxonomic identification quickly.

> If some fish is identified as X, and it's from Y, and X doesn't occur in Y, then I would say, "Okay, well that's wrong"...But it's all based on experience, just focusing on a particular group of fish for the last six or seven years (Zoologist 08).

Zoologists also mentioned several occasions when they needed to access physical specimens via museum loans or visits. Some were breaking new research ground or studying specific parts of a specimen (e.g., fins, skeletons, teeth) and needed data that had not been widely documented across museums. Others did not trust the data they were getting. Some examined the physical specimen to verify species identification. Zoologist 19 explained, they *"look at it with an expert's eye in order to verify that the taxonomic ID of the material was correct."* Others, like Zoologist 06, collected his own morphometric data because he was concerned about getting accurate, consistent measurement data across museums, rather than relying on data produced by different people who used different calipers and measurement techniques.

> I can't go back and make sure that the locality is correct, I can't go back and make sure that the preservation process was correct. But what I can do is [know] the [morphological measurements] data I collect is accurate (Zoologist 06).

When requesting museum loans or visits, zoologists needed additional specimen information, such as museum, lot, and specimen numbers, and the quantity of lots and specimens. Identification numbers were useful for requesting specimens and locating them within a museum, while lot and specimen quantities determined whether a loan or visit would be worthwhile, given the size of the collection. Information about the condition of specimens also was of interest, but as Zoologist 33 explained, for his research on fish fins, it was rarely available.

> You're trying to see where the fins were but they're no longer there…this happens with deep dwelling fishes where you bring them up and they're all like exploded...so that's an information that would have been useful but it's never available (Zoologist 33).

Archaeologists reusing artifact data also wanted to know identification and location information and age of the artifact. Location information went beyond latitude and longitude to include positioning in an archaeological site. Given the destructive nature of archaeological research, the artifact information archaeologists needed often came from descriptions, photographs, and field notes recorded during the excavation. Asked to describe what she meant by "where in the cave something is from", Archaeologist 06 explained the importance of a clear stratigraphy.

> I actually mean what strata it's from. I was talking about the importance of having a clear stratigraphy. And so, if they had labeled stratigraphy, let's say, A, B, C, D, E, and if they're comparing the fauna from E to A, that tells me that when they excavated, they were really careful about preserving that information (Archaeologist 06).

In addition to information about the artifact of interest, archaeologists needed information about other artifacts found with it and their stratigraphic relationship to one another. This information documented the chronological order in which the artifacts were created/deposited in the field and supported the archaeologist's interpretation that occurred in the field and later in the lab. Yet, the information was not always available. Several archaeologists described the various ways they went about getting it, including museum visits and meetings with staff. Archaeologist 02 recalled contacting the excavator for the information because the artifact was separated from the data and he did not have access to the field notes.

> So I was contacting him for other specific information. Where was this found, what period did it date to, and what artifacts were found with it because that's often not cataloged along this primary zooarchaeological data nor do you have access to field notes or anything like that (Archaeologist 02).

The depth of information about the chronology and interconnectedness of the artifacts was important to archaeologists, because this evidence was the only way they could verify the excavators' conclusions in the field. As Archaeologist 18 explained, "I can't go back to the…their [the excavators'] original dispositions because excavated that context has been, that information is gone." Excavation destroys the context of the site apart from what the excavation team recorded.

4.1.3 Data Producer. Data producer information, such as the names of the data creators, their institutional affiliations, and where they were educated, were used across disciplines. Almost two-thirds of the social scientists (63%), more than half of the zoologists (55%), and 14% of the archaeologists mentioned it. While all social scientists used data producer information to assess data quality, novice social scientists mentioned it twice as often as experts (82% vs. 43% respectively). It may be that novices are less willing to base reuse decisions solely on their own evaluation of the data. There also were differences when comparing the zoologists interviewed vs.

observed (70% vs. 23% respectively). The differences may be due to observed zoologists' richer information environment. Their data quality judgments were informed by handling the specimens, not just metadata about the specimens.

In this series of quotes, Social scientist 22 described trusting data producers' proficiency given the institution where they worked, Zoologist 11 discussed the role of a researcher's reputation, and Archaeologist 13 noted differences in excavation based on where researchers were trained.

> And the affiliation of people, the proficiency of the people who are collecting the data, I think it's the most important. In a sense from affiliations. If a dataset had been collected, had been created in some, say, some sort of institution. Well, of course you know the people who work there, they spend some time doing, they have some knowledge or expertise about it, they know how to code, what to code, where to code what they do. So I think that's most important (Social scientist 22).

> A lot of that data …[w]as collected by a single man…who is a quite well-known and well-respected…So, it's a combination of the fact that I knew he knew what he was doing, in addition to the fact that reading through his notebooks, it became very clear, essentially, in the diary aspect of them, that he knew what he was doing (Zoologist 11).

> So there is this whole context issue of... who excavates well. So again there is these other traditions of excavation, whether it's a Turkish... Turks do things differently than those trained in Germany versus those trained in France versus those trained in the US (Archaeologist 13).

Although data reusers relied on personal characteristics as signs that data were likely credible and trustworthy, data producer information was not the only factor considered. Archaeologist 13 explained, "So even if data is from a good zooarchaeologist you trust, the archaeological context might be a little questionable." Study participants across the disciplines mentioned using several types of context to determine data's quality, credibility, and trustworthiness.

4.1.4. Data Analysis. Social scientists mentioned data analysis information most frequently (65%), followed by zoologists (25%) and archaeologists (14%). Most of the social scientists discussing data analysis information were novices (68%). Novices repeatedly mentioned using this information to help them interpret the data, particularly at the variable level. Social scientist 01, a novice, explained.

> So you could have a variable labeled X, but the meaning of X could differ from discipline by discipline. It means the computations or this sort of the scheme that surrounds that word has... Basically, it has different meanings and it's used differently, and it was hard to know how exactly this variable was being used or what it means (Social scientist 01).

Many researchers discussed the need for information about the codes and coding procedures used to transform qualitative data into quantitative data. Social scientist 12 wanted to know how data producers read and applied policy codes to newspaper articles describing House vs. Senate hearings in the United States. Archaeologist 17 described the importance of having access to both the codes and coding procedures when trying to interpret data about soil types.

> There was the issue of a transformation...It was a series of polygons of different soil types, but the classification of the soil types wasn't really documented either...trying to find out, A, what those alpha-numeric classifications meant, and then, B, what that actually was…but then it was also describing soil types based on, sort of, strictly chemical terms or whatever or geological terms, which didn't really have any explanations about them either… (Archaeologist 17).

Data analysis information also included descriptions of how data were aggregated or weighted. This information helped social scientists interpret the data and the data producers' results and perform their analysis and interpret their results. Social scientist 09 described how different calculations of a quality of life measure resulted in different interpretations, and Social scientist 33 wondered whether information was available to help aggregate and analyze data across neighborhoods at the same geographic unit as the Census Bureau.

> …if I am measuring quality of life in Indonesia and I say 50% of that variable's value comes from, political freedom, I have to take that into account that this is a quality of life measure that really heavily values political freedom. When I get very different results, if I'm looking at quality of life statistics where its aggregation comes from 50% of the value comes from, "Do you have an enough food to eat for the day?" In a different country, it might result very differently depending on what those values truly represent (Social scientist 09).

> Is there some type of cross-walked file so that I can take information that's for neighborhoods…and can I aggregate this up to a census place level? Or is there some easy way of combining this across for different census tract? Some type of geographic unit… that the census bureau actually recognizes... (Social scientist 33).

Some social scientists described using basic descriptive statistics, such as frequency distributions, standard deviations, and ranges to identify data issues. Others mentioned wanting information about how data were cleaned and how to replicate analyses, but these types of data analysis information were mentioned less frequently and with less specificity.

Zoologists and archaeologists produced radiocarbon date data from specimens and artifacts. Zoologists also produced genetic sequence data. Information about these analytic processes was essential for reuse. Zoologist 05 mentioned reading about a DNA analysis as one way to validate the resulting sequence data. Zoologist 04 described information she needed to understand uncertainty in a radiocarbon chronology so that she could interpret the primary data.

> …if there's a sequence published for one species of something and in the paper they find the species is completely out of place, or the data don't match from what's been found in the past…it could potentially be that the fish was misidentified or there was some contaminant in the sequencing or something…Doesn't mean I wouldn't use, but it would certainly mean that I would want to read about testing or ask for the specimen to confirm the ID or include some of my own of that species…Just to double check that finding (Zoologist 05).

> So if you get a radiocarbon date, depending on the researcher, or depending on the study, there's a lot of other information that would be particularly useful like what kind of material did they get the radiocarbon date from...what lab did you send it to? … How much material would you date in terms of the [batch]? And so, that quality of information is highly variable

> in the database. But it does translate to real uncertainty when you take…the radiocarbon dates and come up with a chronology... so that's a lot of the work I've been focusing on actually is trying to understand uncertainty in the... Not so much in the primary data itself but in this associated chronology which is necessary to interpret the primary data (Zoologist 04).

Despite the different types of data analyses social scientists, archaeologists, and zoologists employed, all needed to know the activities and processes performed on the data during the original analyses, to understand how they impacted reuse.

4.1.5 Missing Data. Social scientists (42%) mentioned missing data information most frequently, followed by zoologists (8%). No archaeologists mentioned them. Although more novice social scientists (50%) mentioned them than experts (33%), there were no discernable differences in how they talked about them. They needed to know what data were missing, how much, why, and how missing data were handled to determine the extent to which missing data might alter their reuse objectives and data analysis

Common reasons cited for missing data from surveys were subjects and data producers' actions. Human subjects dropped out of studies and skipped survey questions. Data producers designed surveys to have skip patterns. Knowing the causes helped social scientists characterize the missing data. Social scientist 38 explained why skip patterns did not result in "true missing data".

> Not everyone ends up being asked the same questions. And sometimes you have to backtrack through the questionnaires and the codebooks in order to figure out who is eligible to be asked and whether the missing data are true missing data or whether the missing data are just due to these skip patterns, this filtering (Social scientist 38).

Data producers also changed or deleted survey questions in longitudinal studies if the question wording was not working well. Social scientist 07 acknowledged this practice, but she did want to track the question's evolution, which she described as a "relational approach".

> So that's really important too that you have a way to know what questions have changed… is there a lot of missing data on this? Have you messed around with the wording?...And if that's the case, then you show me what the related variables are, right? This is part of an experiment. This is the fourth version of the abortion question that you ran. Would you kindly point me to the other variables that I need to be looking at? And so I think that would be useful, too (Social scientist 07).

Social scientists also needed to know how missing data were handled. Some data producers used codes and coding procedures to flag missing data. Social scientist 27 illustrated why this kind of data analysis information was important.

> I mean, even things like missing data codes, that's extremely important for the empirical work…the missing data, for example, PSID it may be 9 for some variables, the data could be 99 for another one, 9,999 for another one. So you have to know what the missing data code is. And then translate that into a dot because the computer might take a 99 as, for example, for the variable education to mean the person had 99 years of education (Social scientist 27).

Social scientist 34 explained that other data producers did something “to fix” missing data and discussed the importance of knowing what they did because undocumented methodological activities limited reuse.

> So if they've done something to fix some of the missing data or if it's just an imputation, you might know what you can pick out and what's imputed or not imputed. You don't always [have] that methodology of how the imputation was done. So that's the thing, do I turn it on, turn it off, am I going to keep it am I not going to keep it (Social scientist 34).

4.1.6 Research Objectives. The social science community (23%) sought data producers’ research objectives most frequently (5 experts and 5 novices). Only 3 archaeologists (7%) and no zoologists mentioned research objectives. The reasons for seeking information about research objectives varied. Social scientists 25 and 28 accessed published literature to understand what data were used to answer research questions. Both social scientists and archaeologists used research objectives to gauge whether data were relevant for their reuse. Archaeologist 02 summarized seeing and assessing information in a journal article and then following up with the article’s author for additional information about the research objectives and other context information.

> If I see a journal article on that I just assess whether or not it's going to be useful for the research questions that I'm asking... but it always, always ends up with me contacting the researcher and asking them, "Well, how did you collect the data? What were your excavation technologies? And then, how did you analyze the data after all this? How did you analyze these animal bones? So, what research question or reference question did you use? (Archaeologist 02).

For Social scientist 32 and Archaeologist 01 the data producers’ research objectives provided a window on what questions could be answered by the data. Using published literature, Social scientist 32 discovered what research was possible with the data before considering his own ideas.

> And with the data I've used, I mean I’ve often been referred to it by seeing it in a publication so I'm aware of at least one or two. But that information which I don’t download but do follow up on is pretty useful. I’ll typically go and download the article and see what were they doing with the data. Again, it gets you before you start your own analysis a sense of what's possible, what's not possible…Typically, I'll go and read the data and methodology section or at least scan them for articles that are recently cited at that dataset (Social scientist 32).

Although the data producers’ research objectives provided context, they also were implicit and confounded with data collection information. Archaeologist 01 discussed the implications of different survey methods used during data collection (e.g. spacing between survey walkers) and the objectives implicit in those methods.

> Sometimes they'll simply declare we were only interested in broad-based information. We were only collecting broad-based artifacts. There's a high concentration in this area, but we disregarded later stuff...We have to look at their field methods and that's for example, did they walk with spacing close enough so that they were picking up dense... Densities of sherd scatters? (Archaeologist 01).

Another proxy for research objectives (an input) were the results (an output). Social scientist 28 described searching for results in publications.

> So you look at the publications, in the journals, in the books, you look at the people that you know that did the data and probably about their background and the kind of work that they published…Well, you're looking to see how they used the data. Are there a lot missing, what kind of constructs they filled, what kind of results they had (Social scientist 28).

Understanding what types of studies were done with data of interest and the types of questions asked was not a frequently mentioned context type, but researchers thought it was useful when evaluating data for reuse and they sought it in publications or from data producers.

4.2 Repository Information

We identified three types of repository information: provenance, repository reputation and history, and curation and digitization. Although all three types were mentioned across the disciplines, provenance was the most frequently mentioned and ranked the second most mentioned context type overall. It also overlapped with other context types. The second most mentioned type of repository information was reputation and history; curation and digitization followed.

4.2.1 Provenance. Zoologists mentioned provenance information the most (55%), followed by archaeologists (50%), and social scientists (44%). Both provenance as source and as traceability were mentioned, although researchers across the disciplines addressed them in different ways. Zoologists were most interested in provenance as traceability. In cases where additional data were created from a physical specimen, such as radiocarbon date and DNA sequence data, information about the lab where the data were created, and the methods used to create the data, were mentioned. This overlaps with some aspects of data analysis information, but was used to build trust rather than understanding in the data. For Zoologist 20 and others, being assured about specimen identification via links from genetic data found in databases to the physical specimens held at museums also built trust in data.

> …they said, 'We collected it [the specimen] in this particular basin, it's this species', but when you go back to other sources that species is not known to occur in that basin…it [the sequence data] was uploaded to GenBank as a particular species. It turns out it is not that species. So, we encounter problems of misidentifications, lack of voucher specimens for which someone could go back and assess or assure that that is what species they're saying it is (Zoologist 20).

Archaeologists looked for provenance as traceability as well. Of interest was keeping dispersed data from the same origin organized intellectually, by ensuring that appropriate linkages were maintained between artifacts and metadata collected from one site. Since archaeological sites are often destroyed during excavation, linking data and artifacts derived from sites through documentation was important. Archaeologist 03 described the process.

> You know, frequently in archaeology, after you document something, you then destroy it. So, the visual documentation usually and the written documentation has to make sense together and they have to make sense separately. And then there has to be a way to make

> sure that whatever visual documentation you created and your written documentations are inseparably linked... (Archaeologist 03).

Social scientists also mentioned traceability, but focused on version control of a dataset or versioning of individual survey questions. Below, Social scientist 42 discussed dealing with different versions of the dataset over time, while Social scientist 07 addressed changes in question wording. The latter was discussed earlier in relation to missing data information. Tracing when, how, and why data went missing served to create the transparency necessary for informed reuse.

> So that's the original data. Then we got some additional data from the US Department of Agriculture Economic Research Service that had some more variables… and then the more recent data are issued by the National Agriculture Statistical Service...And when we got it, it was an ASCII dataset with a printed codebook. Today you would get it as a SAS dataset or an SPSS dataset or a Stata dataset or an Excel Spreadsheet. There is a lot of flavors that it comes in…we convert them to SAS datasets and keep them that way, that's what's convenient for us. And then we have our own internal project documentation which is very detailed. It shows where every variable came from and how we constructed it (Social scientist 42).

> I know it's Version 4, but I don't know why...they [the data producers] could go, 'Look,' it's like a log saying, 'We noticed that...there was a problem with blah-blah-blah…And we fixed this in variable blah-blah-blah. And this is Version 4…' Just sort of transparency in how you deal with your data, I think I would prefer to have that type of information (Social scientist 07).

Researchers expressed provenance as source in terms of the repository and its curation practices. For example, Social scientist 25 spoke of trusting data given data's source.

> I mean, they are in the business of global developments, so you would trust that they're committed to providing high quality data. So due to the reputation as a leading global development institution, you assume that trust, you assume that credibility follows over into their data (Social scientist 25).

Archaeologist 06 discussed the impact curation practices had on her ability to reuse data.

> One of the real issues when working with small fauna is what was the agent...Who's the agent of accumulation? So, in an archaeological site, typically, you'd assume that humans were the ones that were accumulating all these deposits,…even if all this great information about where in the site the deposits came from, even if that's all in the report, if it has been curated in such a way that that information has been lost, or it's likely to have been lost, then that's a no go (Archaeologist 06).

4.2.2 Repository Reputation and History. Researchers across the disciplines mentioned repository reputation and history. Social scientists (42%) and zoologists (33%) mentioned it the most, followed by archaeologists (18%). The repository's institutional affiliation, extent to which it was well known and widely used, collection size, staff size, expertise, and quality of curation and digitization work were used to assess its reputation and decide whether the data were credible and could be trusted. Much of this information was generated and maintained within the disciplinary community through word of mouth and experience searching for and reusing data from repositories. It was not something that was captured and displayed during data discovery. Social scientist 23 and Zoologist 15 exemplified these practices when talking about repositories' "brand name" and "Louisiana fishes".

> It's one of the first social science quantitative data repositories that existed and as far as I know the biggest…obviously it's good to have some sort of brand name or other types of things (Social scientist 23).

> Well, it's like I said, the two biggest sources came from our collection, and the collection in Monroe which are both major repositories for Louisiana fishes and have had experts in those collections that have long standing histories and are well-respected (Zoologist 15).

Although repository information helped assess data, the information was not about the data, so assessment could only go so far. This may be why the percentage of researchers who mentioned it was lower across the disciplines. Archaeologist 02 explained, "I don't give [the archaeological repository] a sort of blanket trust that all the data in there is correct…they provide enough metadata for me to check that on my own…I sort of trust going there because I know that I can find the information I need to validate it." In other words, the repository has developed a reputation and history around providing the necessary information about data that helps researchers determine data's reusability.

4.2.3 Curation and Digitization. Curation and digitization information was mentioned by zoologists primarily (43%), followed by archaeologists (14%) and social scientists (9%). No strong, overarching patterns emerged about the need for curation and digitization information. Like Zoologist 13, a couple of zoologists interested in tissue samples mentioned museum preservation practices, given the differing effects of formalin and ethanol on specimens.

> Historically, museums specimens went into formalin and then into alcohol. Well, the exposure of the tissue to formalin does a real number on DNA, degrades it very rapidly… makes it much harder to get usable DNA (Zoologist 13).

A few zoologists mentioned wanting documentation about data limitations. Zoologist 09 saw this as one way to enable more rigorous, accurate data reuse.

> Knowing that [the weaknesses of data or data points] would allow people to make decisions of whether to include that point or not include that point in their study… it would facilitate enormously…a critical or a more accurate use of the data (Zoologist 09).

As for error correction, some zoologists expressed surprise that there were errors in museum collections, whereas others expected it. This is evidenced in quotations from Zoologists 17 talked about errors in location data.

> I was amazed that I had specimens of coelacanth that when I mapped them it came out in the middle of the Amazon (Zoologist 17).

Zoologist 34 explained that errors varied from university to university depending on whether museum staff verified the data upon deposit, and zoologists recognized and appreciated curation processes that added value. Zoologist 12 explained GenBank's curation process, which included cross-validating a sequence upon accessioning.

> They will actually BLAST it [the sequence], so compare it to other sequences that are available for that gene and if something's wildly wrong, they will spot it (Zoologist 34).

Few social scientists and archaeologists spoke about curation information. Beyond general statements about its importance, few mentioned the value of social science repository staff who documented, processed, and provided "clean" (i.e. non-sensitive) data for reuse. Interestingly, all archaeologists discussed information they needed about the data, but did not link it to museum or repository processes.

Digitization was mentioned less than curation. Four zoologists and one archaeologist mentioned two aspects of digitization: 1) information about the digitization process and 2) information about standard digitization procedures in place. Speaking about spatial data that had been digitized, Archaeologist 17 described basic information not typically available "where it's been digitized from, when it was digitized, how it was digitized." Two zoologists mentioned needing to know the origin of digitized data, which supports provenance as traceability. Zoologist 04 suggested flagging different levels of data quality depending on data's origin.

> I believe they…actually scanned the pollen diagrams from a paper and got out count data or relative abundance data that way, so you are not even getting original data from the researcher. You're relying on the scanned copy of a figure in a paper. And one thing that is concerning and I think is a general issue faced by the data users and database managers is how to indicate, how to flag these items of differing quality (Zoologist 04).

Zoologist 14 and Zoologists 36 discussed the importance of two standard digitization procedures. Zoologist 14 described the data format standards used when digitizing museum collections that helped to easily combine data across museums.

> So, anybody who decides to digitize a museum collection knows that 'in order to deposit or share my information with GBIF or MaNIS or ORNIS, I have to follow this format.' So that's nice because…if I download data for a species that it comes from five different museums, those data are all in the same format (Zoologist 14).

Zoologist 36 discussed the possibilities of standard procedures to create consistent photographed specimen measurements across museums, including trained photographers and proper orientation and placement of the scale bar.

> If there was like one or two people that their job was to document all of the specimens photographically and they were consistent doing it all the time, then I would probably be more trusting…(Zoologist 36).

Provenance was the second most mentioned type of context, perhaps because it often overlapped with other context types. Provenance as traceability not only referenced the initial origin of physical and digital data, but also when, where, and how the data were transformed through data analysis processes. Likewise, provenance as source was tightly intertwined with data producer, repository reputation and history, and curation and digitization information.

4.3 Data Reuse Information

Data reuse information, the third context category, was comprised of three context types: prior reuse, advice on reuse, and terms of use. Social scientists led with the most mentions in each category, zoologists followed. Archaeologists did not mention any of these three categories, with the exception of one archaeologist who mentioned advice on reuse. This is not surprising, given that archaeological data sharing and reuse, by comparison, is still a recent phenomenon. Although among the least mentioned context types, findings showed data reuse information played a key role in reuse decisions.

4.3.1 Prior Reuse. Prior reuse information was primarily published data reuse studies. Only social scientists (58%) and zoologists (23%) mentioned prior reuse information. Although novice and expert social scientists' reasoning for needing prior reuse information was similar, more novices (73%) mentioned it than experts (43%). Researchers used publications to determine the extent to which the data were reused. Data that appeared in publications indicated a level of peer review and acceptance within a disciplinary community. Social scientist 37 discussed the role of prior reuse as a collective judgment when formulating her opinion about trusting data.

> All of my data reuse experience, all of the secondary data that I used is federally funded, is nationally representative, and is very widely used among the fields that I'm working in so I've never had to make like an independent judgment of, 'Do I trust this survey or not?' It's more a question of this is the survey that is carried out by an entity that I trust and that I know other people that I work with trust and so I just rely on the collective judgment to say that this a trustworthy survey (Social scientist 37).

Social scientists also used published studies to learn what other reusers did with the data and whether they had critiques or encountered problems or limitations with the data during reuse. Social scientists used this information to vet data, as well as to check whether their research questions were novel, or their results would be comparable across studies. Fewer zoologists mentioned prior reuse when making reuse decisions. Those that did spoke of vetting and peer acceptance of data and repositories. Zoologist 34 discussed relying on publications of previous uses of Michigan wolf data from museum collections to make sense of her data and validate specimen identification.

> And there have been people I know who have come through previously, trying to sort dogs and wolves, and what did they think? Maybe some of these animals that we have are actually dogs, and that's making our data noisy. So I would go back and look at whatever publications anyone's done on wolf dogs in Michigan (Zoologist 34).

4.3.2 Advice on Reuse. Social scientists mentioned advice on reuse the most (40%). Only two zoologists (5%) and one archaeologist (5%) mentioned seeking advice on reuse. Although the numbers were approximately equal for expert and novice social scientists, there were some differences in how they received the advice. Both mentioned getting advice about working with datasets through workshops and courses. However, expert social scientists also expressed appreciation for documentation that included recommendations for working with the data (e.g. combining, weighting, or linking data, etc.). Documentation describing how to create indexes and scales was helpful and saved time for Social scientist 38.

> Some of the documentation that's available explains to you as a user actually how you can create indexes and scales from a number of the items that are available which is very helpful, it saves you a lot of time, for example (Social scientist 38).

For novice social scientists, advice on working with the data came from professors and advisors more frequently than documentation. Social scientist 17 described consulting with her advisor to determine how to deal with a variable when the questions kept changing.

> Well the question wording differs between years and so when I'm trying to get it at a complex variable like political identity. If they're not asking the same question over years…are then people answering differently and so there were several discussions that I had with my dissertation advisor about, "Can we still use these if they changed the question wording?" "Is it different or the same enough such that it either should be thrown out or it should stay in?" (Social scientist 17).

Zoologist 11 talked with a colleague, because he was going to collect bird survey data and wanted to combine it with data collected from the same location in the 1940s and 1980s. The colleague had collected the data in the1980s and he had talked with the person who collected the data in the 1940s. Supervisors advised Zoologist 14 about what characteristics of the Seychelles Caecilians to measure and closely examine during her museum visit. When thinking about the analysis she wanted to perform, Archaeologist 9 mentioned sending the codebook she created for the data she was reusing to five other archaeologists who had worked in similar circumstances.

4.3.3 Terms of Use. Although no archaeologists mentioned terms of use information, an approximately equal percentage of social scientists (12%) and zoologists (10%) did. Expert social scientists discussed the difficulties encountered trying to gain access to restricted data and maintain confidentiality. Social scientist 33 explained his frustrations with "double secret restricted access", which amounted to requesting access to restricted data twice, once for the dataset and a second time for certain variables within the restricted dataset. Social scientist 44 discussed the need for more transparency - "to know what's restricted and what's not restricted and then how to launch those procedures for getting that would be very helpful." Integrating data from multiple datasets, Social scientist 31 described the difficulty associated with negotiating confidentiality restrictions.

> They're written in different ways, require different protection mechanisms, and it's very hard to do data integration to negotiate multiple data agreements. And we really need some work on standardizing terms of use for confidentiality across usage data (Social scientist 31).

Actions taken to protect confidentiality were also mentioned. In one case, rather than terms of use, Social scientist 42 was concerned about how data producers' desire to protect proprietary information influenced how they reported the data they made available for reuse.

> How you report on small cell sizes in tabular data is a really important methodological question. Do you completely blank out the ostrich's column or row actually in the table or do you put in there and put an asterisk in it saying there are some ostrich's in Washtenaw County, but we won't tell you how many farms have them, and we won't tell you how many

> they have and we won't tell you how much value they have...It's important to understand how that's done (Social scientist 42).

By examining context from the data reuser's point of view within three disciplines, we identified twelve context types needed when deciding whether to reuse data. Interestingly, reusers' needs extended beyond data production information to include information about the repository and data reuse. Moreover, findings showed several context types were tightly intertwined within and across the three context categories. Some built on one another, while others were brought together to provide a more complete picture of events. Together, these findings have implications for research and practice as discussed below.

5. Discussion

Digital curation researchers primarily focus on preservation activities to make content renderable over time (Lee, 2011; Beaudoin, 2012b). This study expanded the discussion to curation activities that enable meaning making over the long term by preserving data's context as well as content. With a specific focus on data reusers, this study examined the context researchers needed to decide whether to reuse data. Our aim was to identify the various context types data producers and repository staff ought to capture and curate. Considering context within a Collaboratory, Chin and Lansing (2004) identified several context types needed for collaborative data sharing and reuse among biologists. Drawing from their research, we examined social scientists, zoologists, and archaeologists' data reuse practices via repositories.

Although similar in spirit to Chin and Lansing's description of community collaboration, where researchers share their data with their disciplinary communities via a data portal outside of the Collaboratory, our study offers a more nuanced view of a community's reuse needs. When it comes to data reuse in absence of collaboration, Chin and Lansing (2004, p. 416) suggest, "Together, data provenance and annotation features provide researchers some initial capacity to assess the quality of collected data." While we found provenance offered some initial level of trust, ultimately researchers demanded much more before reusing the data. They required the same richly detailed context information that Chin and Lansing provided for BSC members, in addition to some new context categories and types given they were not engaged in collaborative data sharing and reuse. Together these findings offer several contributions.

First, we found that social science, zoological, and archeological communities relied on four key context types Chin and Lansing (2004) identified: 1) general data set properties, 2) experimental properties, 3) data provenance, and 4) analysis and interpretation. This suggests some degree of generalizability with respect to context needs across disciplines and data sharing and reuse environments (e.g. collaborative sharing and reuse via a Collaboratory vs. independent data sharing and reuse via a repository). However, our findings also showed the need for terminology that is more broadly applicable across multiple disciplines and encompasses stated needs. For instance, findings indicated data collection information to be more appropriate than experimental properties, given the range of research methods used across the disciplines examined.

Second, findings indicated some of Chin and Lansing's (2004) context types were more prominent than others. From the general data set properties, researchers in our study mentioned data owners most frequently, which according to Chin and Lansing (2004) are the people responsible for creating the data. This confirms prior data reuse studies, where information about the person who created the data was one of several criteria used to establish trust in the data (Jirotka et al., 2005; Zimmerman, 2008; Van House, 2002; Van House et al., 1998). However, our findings showed it useful to distinguish and retain context information about the entity responsible for creating the data (i.e. data producer) from the entity that legally owned the data (i.e. data owner). In some cases, data producers were evaluated, but in other cases, their institutions had reputations and history, which served as a proxy for the individuals creating data (e.g. The World Bank, U.S. Government, etc.).

In contrast, the other context types Chin and Lansing (2004) identified as part of the general data set properties category (e.g. filename, file type, file size, creation date and time, etc.) were not frequently mentioned in this study, but that does not mean they should be abandoned. This study primarily considered context information needed to evaluate data for reuse. Prior research shows data reuse is a process that includes several other stages, including discovery, access, selection, preparation, and analysis (Faniel et al., 2012; Rolland and Lee, 2013; Zimmerman, 2007, 2008). Different types of context may be important at different points in the process. Once evaluation shows the data have potential and the researcher wants to access data for further exploration, filename, file size, and data format may become more important. Different types of context also may be important depending on the type of data reuse study being conducted. If the goal is showing that research is reproducible, then information about when data were created and modified (i.e. date and time stamps) becomes important.

Third, findings identified three new types of context not found in Chin and Lansing's (2004) study: 1) information about specimens, 2) artifacts, and 3) missing data, which were specific to a particular discipline. This suggests repository staff that steward data across multiple disciplines (e.g. institutional repositories, etc.) not reduce curation efforts to only capturing common context types across multiple disciplines. Instead, it is essential that staff mindfully monitor the unique needs of the communities they support through regular formal and informal feedback mechanisms, otherwise key context information might not get captured and deposited. Consider the differences in specimen and artifact information. Zoologists relied on two commonly collected pieces of specimen information when evaluating data for reuse: specimen identifiers and geolocation. Any specimen data beyond identifiers and geolocation (e.g., measurement data, teeth counts) tended not to be trusted given different, unknown approaches to capturing the data. In these cases, zoologists collected their own data from the physical specimens.

In contrast, archaeologists wanted rich artifact data beyond artifact identifiers and geolocation to evaluate data for reuse. They sought the context information excavators captured in situ when making interpretations and creating data was needed (e.g. stratigraphy and related artifacts). Social scientists focused on missing data. Given the nature of quantitative data collection, it was important to know whether missing data were the result of survey design (i.e. intentional question skipping given the answer to a prior survey question or changes to survey instrument over time) or participant actions (i.e. question skipped in mistakenly or intentionally). Social scientists wanted information describing what, how much, and why data were missing to consider the impact on their intended analyses and decision about whether to reuse the data.

Fourth, by examining data reusers working independently from data producers and relying on repositories for data, this study also contributed two new context categories – repository information and data reuse information. Findings showed these independent data reusers needed information about the data steward (i.e. repository information). This included provenance, which Chin and Lansing identified, as well as curation and digitization information, and the repository's institutional and staff reputation and history. In absence of collaboration data reusers, particularly social scientists, also sought information from the disciplinary community via data reuse information. They were particularly interested in data reuse studies found in the published literature, which suggests the importance of generating data reuse metrics (e.g., Ingwersen and Chavan, 2011; Fear, 2013) and linking these studies to the data reused via digital object identifiers (DOIs) and data citation guidelines (Faniel and Yakel, 2017). Advice on reuse was also important, and here novices primarily relied on advisors, disciplinary experts, and peer-reviewed literature reusing the data. Although terms of use were not mentioned as frequently, findings suggest that reusers would benefit from terms of use that increase transparency about how to access restricted data and consider how different terms of use agreements for different datasets impact data integration and whether and how the multiple agreements can be more easily negotiated.

Our findings have several implications for repository management. First, all types of data reusers wanted richly contextualized and documented data. There was widespread agreement in the need for several types of data production information. Second, disciplines do have distinct differences in the types of context information desired. While every repository cannot cater to every designated community, understanding these differences and the ensuing documentation needs for targeted communities could enhance reuse. Although novices relied on advisors and disciplinary experts, findings showed advice could be documented as well as communicated via workshops and courses. Knowing what reuse advice is valued could be used to improve repository documentation as well as to design services, such as data reuse curricula, especially for disciplinary communities like archaeology, where data sharing and reuse is still a relatively new phenomena. Data reuse metrics might also be used by repository staff to inform data selection decisions and to justify the repository's value in preserving detailed context along with the data, because both are key to sustainability.

When curating some of these new context categories, repository staff must consider how to present the tangible versus intangible outputs needed for data reuse decisions. As an example, many data production and reuse elements can be represented by metadata, whereas repository and staff reputations are by nature more social, less easily codified, and more relative. In discussions of systems design, Dourish (2004) argues that context is a representational and an interactional problem, making solely technical or social solutions alone problematic. Given findings from this study, we also conclude context in data reuse has representational and interactional aspects. Furthermore, reuse itself is an interactional problem. Each reuser and reuser group (i.e.

designated community) has a slightly different set of context needs depending on factors such as their reuse needs, level of expertise, and disciplinary tradition of sharing and reusing data (Bettivia, 2016; Faniel and Yakel, 2017; Faniel et al., 2012).

6. Conclusion

Data reuse requires not only the preservation of data. Rather, data's meaning or the context of data production must also be preserved. Although this may seem to be an obvious conclusion, having data producers and curators put it into practice is difficult in absence of an understanding of data reuse needs. Our findings provide that understanding. We confirmed prior research on the importance of several types of context describing data's production and identified new ones, including several related to repositories and reuse. Together the twelve context types capture data production, repository, and data reuse information that data producers and curators can use to guide the capture and management of data's meaning during data collection and deposit.

References

Akmon, D., Zimmerman, A., Daniels, M., and Hedstrom, M. (2011), "The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs", *Archival Science*, Vol. 11 No. 3-4, pp. 329–348, available at https://doi.org/10.1007/s10502-011-9151-4.

Baker, K.S., and Yarmey, L. (2009), "Data Stewardship: Environmental Data Curation and a Web-of-Repositories". *International Journal of Digital Curation*, Vol. 4 No. 2, pp. 12–27, available at https://doi.org/10.2218/ijdc.v4i2.90.

Beaudoin, J.E. (2012a), "Context and its role in the digital preservation of cultural objects", *D-Lib Magazine*, Vol. 18 No. 11-12, available at https://doi.org/10.1045/november2012-beaudoin1.

Beaudoin, J.E. (2012b), "A framework for contextual metadata used in the digital preservation of cultural objects", *D-Lib Magazine*, Vol. 18 No. 11-12, available at https://doi.org/10.1045/november2012-beaudoin2.

Berg, M., and Goorman, E. (1999), "The contextual nature of medical information", *International Journal of Medical Informatics*, Vol. 56, No. 1-3, pp. 51–60, available at https://doi.org/10.1016/S1386-5056(99)00041-6.

Bettivia, R.S. (2016), "The power of imaginary users: Designated communities in the OAIS reference model", *Proceedings of the Association for Information Science and Technology*, Vol. 53 No. 1, pp. 1–9, available at https://doi.org/10.1002/pra2.2016.14505301038.

Birnholtz, J.P., and Bietz, M. (2003), "Data at work: supporting sharing in science and engineering", in *GROUP '03 Proceedings Of The 2003 International ACM SIGGROUP Conference On Supporting Group Work*. ACM, Sanibel Island, FL, pp. 339–348, available at https://doi.org/10.1145/958160.958215.

Buneman, P., Khanna, S., and Tan, W.-C. (2001), "Why and where: a characterization of data provenance", in *Proceedings from the International Conference on Database Theory*. Springer, London, pp. 316-330.

Carlson, S., and Anderson, B. (2007), “What are data? the many kinds of data and their implications for data re-use”, *Journal of Computer-Mediated Communication*, Vol. 12 No. 2, pp. 635–651, available at https://doi.org/10.1111/j.1083-6101.2007.00342.x.

Chin, G., and Lansing, C.S. (2004), “Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory”, in *CSCW ’04 Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. ACM, Chicago, pp. 409–418, available at https://doi.org/10.1145/1031607.1031677.

Consultative Committee for Space Data Systems (2012), *Space data and information transfer systems — audit and certification of trustworthy digital repositories, standard no. ISO 16363:2012 (CCSDS 652-R-1), Consultative Committee for Space Data Systems, Washington, D.C.*

Darch, P.T., Borgman, C.L., Traweek, S., Cummings, R.L., Wallis, J.C., and Sands, A.E. (2015), “What lies beneath?: knowledge infrastructures in the subseafloor biosphere and beyond”, *International Journal on Digital Libraries*, Vol. 16 No. 1, pp. 61–77, available at https://doi.org/10.1007/s00799-015-0137-3.

Dourish, P. (2004), “What we talk about when we talk about context”, *Personal and Ubiquitous Computing*, Vol. 8 No. 1, pp. 19–30, available at https://doi.org/10.1007/s00779-003-0253-8.

Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C., and Borgman, C.L. (2011), “Science friction: data, metadata, and collaboration”, *Social Studies of Science*, Vol. 41 No. 5, pp. 667–690, available at https://doi.org/10.1177/0306312711413314.

Faniel, I., and Yakel, E. (2017), “Practices do not make perfect: disciplinary data sharing and reuse practices and their implications for repository data curation”, in *Curating Research Data Volume 1: Practical Strategies for Your Digital Repository*, Association of College and Research Libraries Press, Chicago, IL, pp. 103–126.

Faniel, I.M., and Jacobsen, T.E. (2010), “Reusing scientific data: how earthquake engineering researchers assess the reusability of colleagues’ data”, *Computer Supported Cooperative Work*, Vol. 19 No. 3-4, pp. 355–375, available at https://doi.org/10.1007/s10606-010-9117-8.

Faniel, I.M., Kansa, E., Whitcher Kansa, S., Barrera-Gomez, J., and Yakel, E. (2013), “The challenges of digging data: a study of context in archaeological data reuse”, in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, New York, pp. 295–304, available at https://doi.org/10.1145/2467696.2467712.

Faniel, I.M., Kriesberg, A., and Yakel, E. (2012), “Data reuse and sensemaking among novice social scientists”, *Proceedings of the American Society for Information Science and Technology*, Vol. 49 No. 1, pp. 1–10, available at https://doi.org/10.1002/meet.14504901068.

Fear, K., and Donaldson, D.R. (2012), “Provenance and credibility in scientific data repositories”, *Archival Science*, Vol. 12 No. 3, pp. 319–339, available at https://doi.org/10.1007/s10502-012-9172-7.

Fear, K.M. (2013), “Measuring and anticipating the impact of data reuse”, Doctor of Philosophy (Information), University of Michigan, Ann Arbor, MI.

Huang, H. (2015), “Domain knowledge and data quality perceptions in genome curation work”, *Journal of Documentation*, Vol. 71 No. 1, pp. 116–142. https://doi.org/10.1108/JD-08-2013-0104.

Ingwersen, P., and Chavan, V. (2011), “Indicators for the data usage index (dui): an incentive for publishing primary biodiversity data through global information infrastructure”, BMC Bioinformatics, Vol. 12 Supplement 15, available at https://doi.org/10.1186/1471-2105-12-S15-S3.

Inter-university Consortium for Political and Social Research (n.d.), “About ICPSR”, available at https://www.icpsr.umich.edu/icpsrweb/content/membership/about.html (accessed 16 October 2015).

Jirotka, M., Procter, R., Hartswood, M., Slack, R., Simpson, A., Coopmans, C., Hinds, C., and Voss, A. (2005), “Collaboration and trust in healthcare innovation: the eDiaMoND case study”, Computer Supported Cooperative Work Vol. 14 No.4, pp. 369–398, available at https://doi.org/10.1007/s10606-005-9001-0.

Kansa, S.W., Kansa, E.C., and Schultz, J.M. (2007), “An open context for near eastern archaeology”, *Near Eastern Archaeology*, Vol. 70 No. 4, pp. 188–194.

Lee, C.A. (2011), “A framework for contextual information in digital collections”, *Journal of Documentation*, Vol. 67 No. 1, pp. 95–143, available at https://doi.org/10.1108/00220411111105470.

Miles, M.B., and Huberman, A.M. (1994), *Qualitative Data Analysis: An Expanded Sourcebook*. Sage Publications, Thousand Oaks, CA.

Open Context (2015), “About open context: publishing”, available at http://opencontext.org/about/publishing (accessed 29 April 2015).

Pearce-Moses, R. (2005), *A Glossary of Archival and Records Terminology*. Society of American Archivists, Chicago, IL.

Rolland, B., and Lee, C.P. (2013), “Beyond Trust and Reliability: Reusing Data in Collaborative Cancer Epidemiology Research”, in *CSCW ’13 Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, San Antonio, TX, pp. 435–444.

Scott, W.A. (1955). “Reliability of content analysis: the case of nominal scale coding”, *Public Opinion Quarterly*, Vol. 19 No. 3, pp. 321-325, available at https://doi.org/10.1086/266577.

University of Michigan Museum of Zoology (2015), “About the museum”, available at http://www.lsa.umich.edu/ummz/about/default.asp (accessed 20 April 2015).

Van House, N.A. (2002), “Trust and epistemic communities in biodiversity data sharing”, in *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, Portland, OR, pp. 231–239.

Van House, N.A., Butler, M.H., and Schiff, L.R. (1998), “Cooperative knowledge work and practices of trust: sharing environmental planning data sets”, in *CSCW ’98 Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*. ACM, Seattle, WA, pp. 335–343, available at https://doi.org/10.1145/289444.289508.

Wang, R.Y., and Strong, D.M. (1996), “Beyond accuracy: what data quality means to data consumers”, *Journal of Management Information Systems*, Vol. 12 No. 4, pp. 5–33.

Yoon, A. (2016), “Red flags in data: learning from failed data reuse experiences”, in *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives Through Information & Technology*. American Society for Information Science, Silver Springs, MD, pp. 126:1–126:6.

Zimmerman, A. (2007), “Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse”, *International Journal on Digital Libraries*, Vol. 7 No. 1-2, pp. 5–16, available at https://doi.org/10.1007/s00799-007-0015-8.

Zimmerman, A.S. (2008), “New knowledge from old data: the role of standards in the sharing and reuse of ecological data”, *Science, Technology & Human Values*, Vol. 33 No.5, pp. 631–652, available at https://doi.org/10.1177/0162243907306704.

Biographies

Ixchel M. Faniel, PhD is a Senior Research Scientist at OCLC. Her interests include improving how people discover, access, and use and reuse content. Her current research examines how academics manage, share, and reuse research data and librarians' experiences designing and delivering supportive research data services. Her research has been funded by the National Science Foundation, Institute of Museum and Library Services, and National Endowment for the Humanities. Her ORCID is http://orcid.org/0000-0001-7302-5936.

Rebecca D. Frank, PhD, University of Michigan School of Information, conducts research in the areas of digital preservation, digital curation, and data reuse, focusing on social and ethical barriers that limit or prevent the preservation, sharing, and reuse of digital information. Her current research examines the social construction of risk in the audit and certification of trustworthy digital repositories. Her work has been supported by the National Science Foundation and the Australian Academy of Science. Her ORCID is: https://orcid.org/0000-0003-2064-5140.

Elizabeth Yakel, PhD, is a Professor at the University of Michigan School of Information, where she teaches in the archives and records management and digital preservation areas. Her research focuses on users of primary sources, particularly how to facilitate access to digital archives and the reuse of research data. She is currently working on an IMLS-funded project, "Qualitative Data Reuse: Records of Practice in Educational Research and Teacher Development," which examines data reuse by researchers and teacher-educators of digital video records of practice (http://qualitativedatareuse.org). Her ORCID is http://orcid.org/0000-0002-8792-6900.

Acknowledgements

The "Dissemination Information Packages for Information Reuse" (DIPIR) Project was made possible in part by a National Leadership Grant from the Institute of Museum and Library Services, LG-06-10-0140-10, with additional support from OCLC and University of Michigan. We thank members of the DIPIR team, including University of Michigan students, research fellows, institutional partners, and individual collaborators.