Managing Fixity and Fluidity in Data Repositories

Morgan Daniels
University of Michigan
School of Information
105 S. State Street
Ann Arbor, MI 48109-1285
+1 (734) 222-1769
mgdaniel@umich.edu

Ixchel Faniel
OCLC
OCLC Research
6565 Kilgour Place
Dublin, OH 43017-3395
+1 (614) 764-4370
fanieli@oclc.org

Kathleen Fear
University of Michigan
School of Information
105 S. State Street
Ann Arbor, MI 48109-1285
+1 (908) 294-1392
kfear@umich.edu

Elizabeth Yakel
University of Michigan
School of Information
105 S. State Street
Ann Arbor, MI 48109-1285
+1 (734) 763-3569
yakel@umich.edu

ABSTRACT

Data repositories walk a fine line between the fixity and fluidity of the data they curate. Change is constant, but too much change affects the integrity of data. This paper examines data transformations in three repositories, serving the zoological, archaeological, and quantitative social science research communities. Based on in-depth analysis of 27 interviews, we identify a typology of changes: adding value; correcting errors; creating consistency; changing representations of data to reflect new knowledge; responding to designated communities; and evolving practices around collecting. Then we discuss the nature of these changes in terms of the data and collections. Our findings indicate that organizational differences and the diverse needs of the repositories' designated communities play a large role in how they manage change.

Categories and Subject Descriptors

H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *management*.

General Terms

Human Factors, Theory.

Keywords

Data repositories, data reuse, digital preservation, data curation.

1. INTRODUCTION

Scholars have reflected on and debated the issue of fixity and fluidity of digital information for decades [10, 2]. Brown and Duguid [3] note that these two qualities "pull against one another" (p. 198). This tension is perhaps most apparent in data repositories where data integrity, authenticity, and access are on-going goals and concerns. In this paper we examine this tension through an analysis of the types of changes to data that data repositories must manage. The research question motivating this study is: How do repository staff manage changes to data over time? Our data includes 27 interviews with staff from three data repositories, serving zoological, archaeological, and quantitative social science research communities. Our findings indicate a number of changes affecting the data beginning at data submission and continuing through the archiving and dissemination processes.

This paper builds a typology of changes to data made by repository staff and explores the reasons for similarity and variation in the practices of the three disciplinary repositories.

2. LITERATURE REVIEW

There is an inherent tension in the digital curation literature. Maintaining the authenticity and reliability of data over time is linked to preserving the essence of the data. At the same time, the digital curation process can entail changes to data and the significant properties associated with those data. Some of these changes occur through an intervention by the data producer, others changes are made by the repository and may either be related to the data themselves or their metadata; still others are external and may signal shifts in the practices of one or more designated communities of users, e.g., an identified group of users who work with particular types of data [12, 13]. The important point is that changes are controlled and that changes are documented over time.

The inevitability of changes to data is widely acknowledged in the digital curation community. The Open Archival Information System (OAIS) Reference Model [12, 13] assumes that digital objects will change over time. The UK InSPECT project [11] established a baseline for determining authentic records, and this accommodated the need for change. Building on this work, Coyne et al. [4] assert that designated communities must accept the fact that many digital objects cannot be preserved in their original states and that the original bitstream changes in any migration process. Perhaps nowhere is the inevitability of change more apparent than in the Australian Performance Model [7] which ascribes fixity to the underlying digital object (source) but fluidity to each rendering (process). "When a source is combined with a process, a performance is created and it is this performance that provides meaning to a researcher. When the combination of source and process ends, so does its performance, only to be created anew the next time the source and process are combined" (p. 9).

What is less apparent in the literature is that the impetus for change varies. Repositories enact changes during routine digital preservation regimes [14] to comply with best practice. Yet, changes in collection, description, and delivery are also community-driven [e.g., 5, 6, 8]. How the appropriate nature and extent of these changes is determined is debated in the digital curation literature as well as in other disciplines. For example, Allison et al. [1] focus on the difficulties of maintaining authenticity over time and the critical role of the repository in this process. "To present digital evidence, those assessing it need to be

Copyright is held by the author/owner(s).

iConference 2012, February 7-10, 2012, Toronto, ON, Canada.
ACM 978-1-4503-0782-6/12/02.

able to trust the processes of curation, transformation, and rendering, taking the bitstream from what was actually created (the content) through to what was experienced by those involved at the time and what they are experiencing in assessing the evidence (p. 371). Kelton et al. [9] cite the importance of predictability, which they define as "the stability of the information over time" (p. 370) for trust in that data. Yet, total fixity impedes preservation activities and can also prevent proper contextualization of the data.

Thus, the literature acknowledges that both the fixity and fluidity of digital data are important. Data repositories manage these forces: a lack of balance or too much tension can have consequences for data integrity as well as repository reputation. This paper discusses changes in data and collections and how repositories accommodate them over time.

3. METHODS

3.1 Site Descriptions

In this study we focused on three repositories representing diverse disciplinary communities. The Inter-university Consortium for Political and Social Research (ICPSR) is a data archive for the social sciences, the University of Michigan Museum of Zoology (UMMZ) collects and preserves specimens and data used by zoologists, and Open Context provides a data publication platform for archaeology.

In addition to their respective disciplinary focuses, these three repositories differ in several important ways, including staff resources and characteristics of their data collections. ICPSR, which was established in 1962, currently holds over 500,000 research files and has approximately 116 permanent staff. The repository houses primarily well-structured, digital, quantitative data and adheres to best practices in digital preservation and data archiving. It is an OAIS-compliant archive, and has been awarded the Data Seal of Approval. Data enters ICPSR through several means. Any producer of social science data can choose to submit a dataset via ICPSR's online submission form; in some fields, data deposit with ICPSR is mandated by funders. ICPSR also recruits data from major studies and contracts with several survey organizations and federal agencies to archive their data.

UMMZ was founded in 1895, although the collection of specimens began in 1837. It now has 15 million specimens and a staff of approximately 48 people organized in 6 divisions, each specializing in a different zoological group. UMMZ collects physical specimens along with digital data about those specimens, in addition to audio recordings and images, and archiving practices can vary between divisions within the museum. Data collection at UMMZ is driven by the research interests of the curators, who are faculty members at the university. In the course of their research, they collect specimens, which are accessioned to the museum. This process is the primary way in which UMMZ's collections grow, though they also occasionally add collections that are purchased from other institutions or donated by private collectors. Much of the data about specimens is available online, but UMMZ also lends specimens out to researchers at other institutions.

Open Context is approximately five years old. With the help of five people, it currently provides access to data from 20 archaeological projects. Open Context is not in and of itself a preservation repository at this time, but its partnership with the California Digital Library (CDL) provides it with archiving and backup services. The data in Open Context varies: it accepts a variety of digital data including field notes, site inventories, images, and geographic information system (GIS) data. The data are largely qualitative and unstructured. Like ICPSR, most of the data in Open Context is contributed by researchers external to the repository. Both solicited and unsolicited data contributions are made to Open Context.

ICPSR's designated community consists primarily of researchers—both students and faculty—at its 700 member organizations, which include colleges, universities and research centers in the U.S. and abroad. The users of ICPSR data are quantitative social scientists working in a variety of fields, some of which align directly with ICPSR's 16 topical archives on demography, aging, substance abuse, and mental health, among other areas.

Whereas access to a portion of ICPSR's data holdings is restricted to individuals affiliated with member organizations, Open Context makes its data available free and open to anyone online. While its collections are strongest in zooarchaeology and Near East archaeology, staff at Open Context consider their designated community to include archaeologists working in any area. UMMZ's designated community is comprised of zoology researchers, although increasingly, climate change researchers are using UMMZ's data.

3.2 Interviews

We conducted 27 interviews with staff members from all three repositories. Our interviewees were staff members who work directly with data as it is being submitted, archived and disseminated as well as those who recruit data and shape the collections at each organization. At ICPSR, we spoke with topical archive managers and data processors; at UMMZ, with curators and collection managers; and at Open Context, to editorial and development staff as well as managers at CDL.

The semi-structured interview protocol solicited interviewees' knowledge of the submission, archiving, and dissemination processes at their institutions. Interviews with ICPSR and UMMZ staff were conducted in each interviewee's office or workspace, and the Open Context interviews were done in person and by phone. All interviews were recorded and transcribed for analysis.

3.3 Analysis

The interviews were analyzed using NVivo, a qualitative data analysis software tool. Our team worked collaboratively to develop the code set and two team members coded the interview transcripts. The code set was developed based on the interview protocol and expanded as new codes arose from our ongoing analysis. After training on several test transcripts, the coders reached a reliability of 0.8 using Scott's Pi, a statistic showing high inter-rater reliability for the coding of textual data.

4. FINDINGS

Our analysis of interviews with staff members revealed a number of practices for dealing with change in data repositories. In this section we discuss the six categories of change that we identified in our data: adding value; correcting errors; creating consistency; changing representations of data to reflect new knowledge; responding to designated communities; and evolving practices around collecting. While these categories are applicable to each of the repositories, the ways in which they apply vary a great deal. Our findings indicate that organizational differences and the diverse needs of the repositories' designated communities play a large role in how they manage change. We describe our categories of analysis and apply them to the three repositories in the remainder of this section.

4.1 Adding Value

Adding value represents the activities repositories undertake to make data suitable for secondary analysis. We found that UMMZ, ICPSR, and Open Context add value in various ways, including cleaning, standardizing, and formatting data for their designated communities. They also perform activities that enhance the datasets and provide tools and services supporting data discovery and reuse. For instance, UMMZ does special preparations of its specimens, ICPSR provides online search and analysis tools, and Open Context engages in a peer review process.

At UMMZ, data are derived from animal specimens that researchers collect in the field. There are standard ways specimens are prepared for later analysis, but in some cases UMMZ decides to do special preparations. For example, in ornithology CA02 explained, "Well, the round skins and the skeletons are the ones that are used the most. And so it would only be for specialized uses where you have to look at some other type of preparation." The type of preparation selected is decided, in part, on the basis of what is currently in the collection: "how many fluids do we have of each sex, how many skeletons of each sex […]". Special preparations are also done for unique specimens. "We get stuff from the zoo, for example. And so those we actually usually save as skeletons and flat pelts, because they tend to be exotic species that we may not have very many specimens of, we want to save the feathers, but the skeletons are also valuable because for most kinds of birds, there are fewer skeletons available than round skins. […] And so we're very interested in having good skeletal representation. And so by saving a flat skin, you can have feathers also" (CA02).

The value ICPSR adds is also in how the data are prepared for reuse. Through its automated processes, ICPSR is able to offer a full suite of data formats for download, including SPSS, SAS, Stata, and ASCII files. Another output of the processes is ICPSR's own version of a codebook for each dataset. A codebook is the data documentation users rely on to make sense of a dataset. One thing ICPSR does when creating a codebook is to include survey question text along with each of the variables in the dataset to make it easier for users to follow. ICPSR also prepares datasets to be used with its online data analysis tool and enables variable searches across datasets through its Social Science Variable Database (SSVD). The SSVD, which is also the result of automated processes ICPSR has in place, currently allows users to search across 1.5 million variables from 2,200 studies. According to CB05, ICPSR deals with a fair number of restricted access datasets. In these instances, ICPSR works with the principal investigator (PI) of the study producing the data to provide as much information as possible to people seeking data for reuse. ICPSR has "[…] the ability to make the decision with the PI to allow the codebooks to be viewed publicly so people can see if it's worth their energy to go through the hoops to get the restricted version or not […]." Being able to access the codebook or the variables through SSVD would be useful, because even though restricted datasets go through the same processes as unrestricted ones, they cannot be downloaded without prior approval.

One way Open Context adds value that differs from UMMZ and ICPSR is engaging in a peer review process at the point of submission. In addition to the processing Open Context does for all data contributions, they also give data producers the option for peer review of their data prior to ingest. An editorial board has been formed to oversee the process. According the CC02, the reviewer would get a list of peer review questions "And if it [the dataset] sufficiently surpasses these various questions then we're going to add a little like stamp or star or something to mark that that dataset has gone through this additional level of scrutiny […]." By using peer review, Open Context is able add value without expending a significant amount of its resources.

Both UMMZ and ICPSR are adding value by preparing data so that it can be accessed and reused in different ways during the dissemination process. UMMZ does a variety of specimen preparations based on the state of the current collection and the uniqueness of the specimen. ICPSR provides services and tools so users can access datasets in different formats and understand and search datasets at the level of their individual variables. We contrast this to Open Context, which is also adding value, but doing so during the submission process and off-loading some of the work to others within the designated community.

4.2 Correcting Errors

Our interviews revealed that both staff and users, at different times, are important sources of error detection for the three repositories. Across repositories, the majority of error correction takes place during submission, when staff members notice problems with the datasets they process. In correcting errors, Open Context and ICPSR emphasize interaction with data producers while at UMMZ, error detection may take place long after the submission stage, as it relies on specialized expertise and close interaction with a dataset. At UMMZ, then, changes are often made at the dissemination stage, based on input from users who notice issues with the data.

At ICPSR, error detection is a specific responsibility of data processors. While UMMZ and Open Context staff will correct errors they find, this is not an explicit job duty at those repositories. Accordingly, at ICPSR errors are most commonly found during the processing of a new submission. Staff members are very careful, though, to consult with data producers before making corrections. As one data processor told us, "We note problems in the data and, you know, we let the PI's... [we] tell them we found X, Y, Z. We can't change it. Not unless they direct us" (CB07).

A more nuanced view of error correction to ICPSR datasets emerged from an interview with another data processor, who differentiated between smaller changes, such as typographical errors or misspellings, which she makes without consulting the data producer, and larger changes, such as which variable goes with which data, a change she consults with the data producer about before making. As she explained, "[…] So let's say that in the listing that we get from them, dataset one is supposed to be family data, dataset two is supposed to be child data. And when we're going through and doing our check, it seems very clear to us that the reverse is true, where child is dataset number one. So just letting them know. 'This is something we observed. Would you check and get back with us?'" (CB04). For smaller changes like typographical errors, CB04 corrects them and then notifies the data producer of those changes when the data are ready to be released.

These interviews convey the fine distinctions that go into the decision making surrounding error correction on the part of repository staff. They also demonstrate differences in the amount of interaction with data producers to correct errors depending on the nature of the error, the processor and data producer's communication styles, and the latter's willingness to have changes made to the dataset. As a rule, the data producer is consulted regarding changes to data, but in practice, that consultation takes different forms.

Open Context staff also expressed concern about making changes to datasets. One staff member told us, "For zooarchaeology, which is what I do, yes, I would, if there's taxonomic names that I would make sure are right and spelled correctly. [...] if someone puts all Latin names and then has 'pig,' to make sure to change that to the Latin name of pig […] Once you get past those really accepted ones, I wouldn't change somebody's records except for the clean-up type stuff" (CC02). Familiarity with the specific research genre, then, is an important part of the error correction process for Open Context. This staff member's experience with zooarchaeology gives her a better understanding of data from that field and a greater ability to spot and correct errors in those datasets. While some errors require a certain amount of expertise in a particular subject area, others, like lack of uniformity throughout the dataset, are easier to find and correct by those without a great deal of specific subject expertise and without direct consultation with the data producer.

At UMMZ, expert users occasionally notice discrepancies in databases, for instance, when a species was reportedly collected in a habitat in which it is not generally found. In these cases changes are generally not made unless the data are totally improbable. However, a UMMZ staff member stated that expert users' responses to database content are, "very useful, because we get quite a bit of feedback from people saying, you know, 'Shouldn't this be a different species?' And we'll say, 'Oh yes, it was a mistake in the database, or that name was wrong in the database.'" As CA02 explained, "not everything in a museum collection may be identified correctly. And so you may run into problems if you aren't kind of an expert, or at least very familiar with the species that you're looking at." Problems with species identification that an expert in that species might spot immediately could be completely invisible to non-expert users.

For UMMZ, then, presenting data to a distributed audience of experts helps the museum identify errors and correct them over time. As we will see in section 4.4, however, incorrect identifications of species are not always "errors," they may reflect the changing state of knowledge in a discipline. UMMZ and Open Context rely a great deal on specific expert knowledge for error detection and correction, leading to some error correction during and after the dissemination of datasets, when a larger audience of users with a greater range of specialized expertise accesses them.

4.3 Creating Consistency

ICPSR, UMMZ, and Open Context all make changes to create consistency to ensure that data and documentation are discoverable and have the same look and feel from one search to the next. Similarly, all three repositories create consistency through the implementation of discipline specific standards.

ICPSR uses the Data Documentation Initiative (DDI), a metadata specification for the social and behavioral sciences, to create consistency in its presentation of codebooks across its thematic collections. With the launch of DDI in 2000 and the implementation of various automated and human processes, ICPSR makes the same set of data formats available for download regardless of the format a data producer submits or the thematic collection a dataset belongs to. The same goes for the ICPSR generated codebooks. As CB10 explained, datasets either come in or are converted to "[…] SPSS and then this [automated] process transforms the original dataset in that format into all the formats that we distribute […]. The same automated process also produces the DDI file, which means that it's a file that documents the variables in this standard, DDI. And what happens behind the scenes is, once the DDI is produced it is also automatically converted into PDF and becomes part of the codebook."

In contrast to ICPSR, UMMZ's internal processes are more human-intensive. In addition, consistent descriptions across the different collections are not the priority. Instead, UMMZ coordinates with other institutions to ensure users can search across distributed mammal, amphibian and reptile, fish, and bird databases. As CA05 explains, coordination across institutions has been happening since the 1970s. "[T]he American Society of Mammalogists put together a series of data centers. I think it's been modified or superseded by something called the Darwin Core, and we all more or less…what we have is a subset of the Darwin Core and we participate with […] other collections in the data portal. So our content is available to anybody by going to this portal, along with the content of these other museums" (CA05). The Darwin Core standard, an extension of the Dublin Core standard for biodiversity information, was created to facilitate data sharing and discovery. However, as CA08 explained, there is also a need to change the location data received from data producers. To ensure consistency with other mammal databases, location information is converted to longitude and latitude decimal degrees. "And so we have to check all that data when we put it in. Not everybody uses decimal degrees or there's a lot of people who don't even use latitude or longitude, they just use an anecdotal description of the locality, '5 miles east of Ann Arbor on the Huron River.' So, we have to check that. […] And we've also done that for all the old material that predates the '80s, '90s, and so forth, because […] what they used back then was mostly township and range […]."

Open Context has implemented ArchaeoML to ensure consistency between the various archaeological datasets they receive. ArchaeoML, a generalized data structure standard for archaeological information, was chosen in part because of its simplicity. CC01 discussed other standards that are sophisticated and useful, for instance, CIDOC CRM. However, "[…] to do a full CIDOC implementation, that would also require a lot more metadata that I think would be difficult to actually get from our contributors. So this is just something that was really promoted mainly at big institutional settings where you essentially had datasets that are sort of institutionally curated that have a lot more resources associated with them, and the kinds of things that we actually get are datasets that are developed by individual researchers." ArchaeoML also has a history of development and use in working field databases. By implementing it, along with a data import tool, Open Context is hoping to make contributions easy enough for archaeologists to do themselves.

Although all three repositories are creating consistency, the differences are with whom or what they want to be consistent. ICPSR has focused on creating consistency across its own thematic collections. To date it has had the advantage of dealing with one kind of quantitative data (i.e. survey data) and having resources to create centralized processes and procedures. In contrast, UMMZ is dealing with a large number and variety of datasets that require specialized processing and fewer resources to devote to that work. Consequently it has opted for consistency within a collection, but not throughout the museum as a whole. Lastly, Open Context is a younger and much smaller repository. Therefore it has decided to focus on creating consistency within its own collection, but also consistency with existing disciplinary practices in hopes of off-loading some of the work related to readying datasets for reuse to the data producers.

4.4 Changing Representations of Data to Reflect New Knowledge

A fourth category of change to data held by repositories relates to changes in the state of knowledge in a field. This section reviews ways in which the repositories we studied reflect changes to the meaning of data when the state of knowledge in the designated community changes. This category is particularly relevant to the zoological collections at UMMZ, where the understanding of a given specimen can change during its time at the museum.

At UMMZ, the identification of specimens varies over time, as systematic research advances in a given field of zoology. Systematic zoological research deals with the identification and taxonomic relationships between animal species, including their description, environmental distribution, and evolutionary history. Collected specimens play a major role in this research, providing the basis for comparison of the morphological features of species and their classifications. Because the understanding of species changes, the *meaning* of a dataset, in this case the specimens collected from one location by a researcher and their associated information, may change long after the data were collected.

When a new identification is made for a given specimen at the museum, it undergoes several changes. The specimen is given a new label reflecting its new name and its physical location also changes. Each division within the museum houses its specimens according to their taxonomic designation, grouping them by family, genus, and species, and within species, by geographic origin. If the species designation for a given specimen changes, that update is made in the museum division's database and in the physical arrangement of the collection. The latter change can have a large impact on the storage of the collection, leading to significant rearrangement of specimens. Because of the work involved in reorganizing specimens, this work is only done once a redesignation has gained general acceptance in the field. As CA04 explained, "we're very conservative in our approach, so when somebody does an analysis and publishes it and says, 'Hey, I'm reorganizing this entire group, and this is the new taxonomic scheme,' [...] We usually let it stand the test of time." CA04 went on to explain that he waits for general acceptance by "talking with other people […] seeing the [specimen] names published over and over again."

CA07 contrasted technical errors and scientific errors, highlighting the fluid nature of zoological identifications. While users will sometimes find small problems with a dataset, more frequently feedback about a dataset reflects "not a technical error but scientific error. It's not really error. Science is not an error. In fact, it changes in time. For example, the specific name can be changed later and the relationship with other groups of animals can change over time. So it's not really error" (CA07).

The representation of specimens at UMMZ, then, is an ongoing task. Part of the creation of zoological knowledge deals with changing identifications of specimens, work undertaken by experts in a particular area both inside and outside the museum. The current identification of a specimen, until verified by an expert, is understood to be temporary. Use of the collection reveals possible problems in identification and expert use allows identifications to be verified or updated.

While the understanding of archaeological and quantitative social science datasets held by Open Context and ICPSR also changes as knowledge advances in those fields, subsequent changes to the representation of data was not as salient an issue. At all three repositories, datasets are the raw material for new analyses and interpretations, however, new interpretations of ICPSR and Open Context data do not change the representations of the data themselves. At ICPSR, those interpretations are additive, captured in the bibliography of studies using the repository's datasets. At Open Context, bibliographic information related to the creation and original research use of data is captured, but a bibliography documenting secondary data use has not yet been developed.

4.5 Responding to Designated Communities

Each of the repositories we studied responds to the needs of their designated communities, whether providing new content or developing services to better support disciplinary practice. In each case the repositories adjust their own work practices to meet the developing needs of their users. When DNA testing gained popularity in zoology, UMMZ began to collect tissue samples. ICPSR is currently dealing with the introduction of new data formats and the needs of social scientists doing interdisciplinary research. In its early stages, Open Context has found it necessary to focus more of its effort on supporting its designated community's work practice rather than content needs.

When DNA testing became available to zoological researchers, UMMZ had to develop new specimen preparation, preservation, and loan procedures. As CA04 explained, "Well, we have, over the last couple of decades, been amassing a frozen tissue collection which is now, I don't know, 15,000 vials that we keep in a minus 80 freezer." Other UMMZ divisions also changed how specimens are preserved, so DNA can be extracted later. "You go back 30, 40 years and traditional practice was to store materials. If you're dealing with soft bodied material, [you're] not putting it in ethanol, but putting it in formalin which fixes the tissues better than ethanol. However, you do that and getting DNA from it is incredibly difficult. And so now, we follow the same practice of preserving everything directly in ethanol and not using formalin at all. And hopefully, that's going to be relevant for future use" (CA11). DNA testing has also influenced the loan requests UMMZ receives. According to CA02, "most of the loans that we do now is actually little clips of skin from lung specimens that people are using for their DNA, or the frozen tissues of the same." In turn, this has influenced UMMZ's loan approval process. Researchers wanting to take samples from specimens, "have to prove they can get usable DNA out of the specimens […] that they have the protocols down and they know how to do it so we're not just wasting that specimen" (CA02).

ICPSR is also dealing with demand for new data formats, such as video. "And no one has really tackled it. And it's ripe, right now; we're going to start moving in that direction. But the earliest we would be disseminating video data, and that would be to a closed group of researchers, would be fall of 2012" (CB03). Similar to UMMZ's experience with tissue samples, it is likely that ICPSR will have to make broad changes to its internal processes with the introduction of video data. ICPSR is also dealing with demands from its designated community of social scientists that intersect with the health and environmental science communities. For those conducting this kind of interdisciplinary research, CB01 explains that, "people are [also] expecting to be able to slice and dice the data at the variable level. So I [the user] am looking for these kinds of variables, and they might exist in two or more data files, and pulling them together. The provenance of pulling them together can't be broken; [I] have to know where it came from […]." ICPSR is developing ways to support this kind of activity as well.

In its early stages of planning, Open Context found that its designated community wants simple ways to access content. CC02 explained how the needs of Open Context's designated community of users did not match its original "pie in the sky picture of what people might do with the data." By conducting user experience discussions and feeling out the community informally, Open Context realized archaeologists "don't want to go in necessarily and tag and make it [the repository] all sort of Flickry and share things […] So we decided to move more toward a more formalized publication platform because the need that we see in the community is that people want something that's more professional. They want to have something where you can put your data in and it's citable and its more […] will be valued more as a quality publication type resource […]" To date the use of data in Open Context has been limited. Given this, CC01 wondered "how much do you get ahead of your users and try to promote things to sort of pave the way for a future better practice and how much you sort of follow what your users want right now." For instance, CC01 noted that data format requests from the archaeological community have been very straightforward. They seem to want data stored as comma separated values so that they can work with them in Excel.

A major difference between Open Context and UMMZ or ICPSR is that Open Context and the archaeological community in general has had less experience with data reuse. This may be one of the reasons why we are seeing differences in what the archaeological community is demanding. In contrast, UMMZ and ICPSR's designated communities have been reusing data longer. Therefore they may be more adept at identifying opportunities for data reuse that have implications for the content the repositories should be holding.

4.6 Evolving Practices around Collecting

Staff members from the three repositories we studied reported expanding the kinds of data they collect into new areas. In this section, we examine the mechanisms for change in data collecting at these repositories. We describe the ways in which selection of new collecting areas takes place differently at the three repositories, motivated by the interests of different stakeholders including data producers, funding agencies, and potential users.

At UMMZ, curators and other collectors associated with the museum initiate new kinds of collections. Specimens are added to the museum primarily based on their usefulness to the research programs of these individuals. For example, UMMZ has a large collection of syrinxes, the vocal apparatus of birds, collected originally by a researcher affiliated with the museum who was interested in that feature of birds' anatomy. This researcher's interest led to the establishment of a new collecting area which is still being developed for reuse, although it is not being actively researched by current curators. A staff member explained, "you can dissect [the syrinx] out of the specimen and preserve it in alcohol. […] It's just like any other taxonomic feature that you can look at, so it varies among the species and you can use it to as a character for analysis to look at differences among species. […]." (CA02).

While the interests of curators and other affiliated researchers determine the data collection focus at UMMZ, Open Context receives the majority of its datasets from researchers who are outside the organization. In some cases, repository staff members approach researchers whose work would be a good complement to the collection. These researchers are identified through their publications or through staff members' personal knowledge of current archaeological work. In some cases, researchers who are already aware of Open Context will offer to donate their data. As an Open Context staff member told us, "we get a lot of people who just have very straightforward, 'I have this analysis I did, would you be interested?' And pretty much, we always say, 'Yes, we're interested, and we want to talk about it further'" (CC02). Because archaeological teams can consist of a large number of people with different areas of specialization and degrees of willingness to deposit data, getting agreement to share data from all stakeholders can be difficult. This has an impact on the data Open Context makes available. As the staff member went on to say, "we're trying to go for the low hanging fruit, right now, the projects that are, everyone's on board and everyone's happy to share. And there's not going to be any issues with people who don't want to share the content [...]" (CC02).

Open Context's response to data sharing and intellectual property contrasts with ICPSR, which has technological and procedural mechanisms in place that allow data producers to specify varying levels of access for different categories of users. CB12, while showing us the computer system that tracks data processing at ICPSR, pointed out, "So this is a processing plan, confidentiality review, initial review, deposit review. So every study gets some of these and this all gets tracked here […]" ICPSR's long history as a data repository is part of the reason for the well-developed procedures in place there.

Determining new collecting areas at ICPSR is markedly different from the other repositories. Acquisitions staff, directors of topical areas of the repository, and the ICPSR council make decisions about the collecting emphasis of the organization. One ICPSR staff member described a comprehensive approach to collection development, including thematic areas that change periodically in relation to a slower changing collection development policy. "Currently, we have an interest in mixed methods studies and then we have sort of this prospective technique to go out and try and cull a good list of mixed methods studies and then go after them, both from the leads database and from other ways. We look at journals and do expert interviews with faculty members that are sort of known mixed methodologists and ask them what they think is good and worthwhile data" (CB15).

In contrast with UMMZ and Open Context, then, changes in collecting emphasis are actively established, researched, and pursued by ICPSR staff. They are also, to some degree, a result of relationships with funding agencies at ICPSR. Funding organizations will often contract with ICPSR's topical archives to collect data from specific studies or research areas. This has an influence on the selection of a portion of datasets ICPSR holds.

The various approaches taken by these repositories in identifying and pursuing new collecting areas reflect their different relationships with data producers and designated communities. While UMMZ determines collecting areas based on the field work and research needs of its curators, an internal user community, ICPSR is more externally focused, seeking out datasets in areas of interest. As a younger organization, Open Context sits somewhere between the two, accepting offered datasets when feasible and seeking out easily shared datasets within the discipline.

5. DISCUSSION

In this section, we analyze broader commonalities and differences between the three repositories, beginning with the ways in which each repository documents the changes they make. Next we discuss the people who act as sources of change for the repositories: primarily repository staff and designated community members. We assert that change to these data repositories, to a great extent, reflects repository relationships with their designated communities.

5.1 Documenting Change

One strategy used to manage change in all three repositories was to document it. By affixing evidence of change, these repositories work to ensure data integrity. While not a type of change itself, we found variation to be representative of the organizations' commitments to accountability and the audiences to which they are accountable. All three repositories have methods for preserving information about changes to data, but with important differences in which changes they document and the how the documentation is used. While ICPSR and Open Context capture a great deal of information about change, they largely use that information internally. UMMZ records less information about change, but makes it available to a wider audience.

ICPSR has an automated method for tracking changes they make to datasets during processing, whether to add value to the data or for the purposes of error correction. The repository's in-house software both manages the submission and archiving processes and records actions taken by staff to complete those processes. While the information recorded by this system is extensive, it is not for public consumption – rather it is kept internally. A notable place where changes are documented for users, however, is in the codebooks processed by ICPSR staff for each dataset. In the case of data collections from longitudinal studies, which are augmented periodically with recently collected data, codebooks record changes in data collected from one iteration of a survey to the next. This information is intended to be used for data evaluation and analysis, if, for example, a researcher wants to compare similar questions asked in a given survey over time.

Open Context is also careful to document changes to datasets over time, however, in their case this means keeping a dark archive of datasets in prior states when new updates are made. Although these archived datasets are not readily available to the public, they are kept in order to make it possible to evaluate new analyses of the data. When citations are made to Open Context data, a specific edition of the dataset is included in the citation, in order to make verification of the data easier. This feature helps reconcile the fluidity of digital data with the need for stability required to understand an analysis of a dataset. A reader may need to know which version of a dataset was analyzed.

UMMZ also documents changes to data by retaining old labels with specimens while including new labels updating an identification and giving the name of the person making the new identification. While this kind of knowledge, showing whose expertise a new identification is based on, is essential to museum staff to document thoroughly, this is the only kind of change to datasets that UMMZ staff members always record. Changes clarifying the location in which a specimen was collected are sometimes made to a database when specimens are accessioned, based on verification with field notebooks or specimen collectors, without necessarily documenting the source.

Differences between the three repositories in documenting change illustrate their approaches to accountability to users and producers of data. While ICPSR and Open Context consult with data producers in making the majority of changes to data and document those changes thoroughly, they keep that documentation largely internal, for their own use, and provide data producers with overviews of the changes they make. In contrast, UMMZ only documents some changes to data, such as the identification of a specimen, while changes to the representation of the location in which a specimen was collected, for example, are made without documenting them. This difference may be due, in part, to the relationship of the repository to their contributing data producers. While data producers are external to ICPSR and Open Context, they are primarily internal to UMMZ. As faculty at the university (and often as curators at the museum), UMMZ data producers share an institutional background with the repository that may build an increased level of trust and alleviate the need to communicate more detailed changes. They also deposit their field notebooks at the museum, which serve as the primary documentation of the specimens collected, allowing for material to be checked against those records rather than requiring direct input from data producers.

These variations in documentation suggest communicative priorities for the three repositories. ICPSR and Open Context's focus on documenting change to datasets shows a concern with accountability to data producers. Even though most of that documentation is used internally by staff, their automated systems for either capturing all changes made in processing or for automatically archiving data sets when corrected versions are released show that both repositories wish to keep evidence of the work they have done. In contrast, UMMZ only documents changes to species identification, implying that identifications and accountability for who identified which specimens are the most important changes for the museum to track.

5.2 Instigating Change

The changes to data and collections discussed throughout this paper have had several major sources: repository staff and designated communities (acting as both data producers and users). At all three organizations, repository staff, responsible for ingest, archiving, and dissemination of datasets, do the day-to-day work needed to make data accessible and useful. They build and implement systems and processes to make data consistent and usable, standardizing metadata, examining datasets for errors, and adding value to the data by providing it in various formats and venues. They negotiate with data producers to gain access to datasets and steer the expansion of collections into new areas.

While staff members perform the bulk of this ongoing work, they do so in response to broader changes in the designated communities they serve. Not only do members of designated communities collect, donate, and use the datasets available in these repositories, but their work practices also guide changes in the kinds of data and data analysis tools that repositories make available. At all three repositories, the types of data provided are changing in response to designated community practices and needs. While UMMZ now preserves specimens in ways that make DNA extraction possible, ICPSR is working to make qualitative data a meaningful part of its offerings. Open Context, in response to a recognized need among archaeologists for a data publication platform, is growing to fill that need, including a mechanism for peer review of datasets. Responsiveness to a designated community is a trait shared by all of these repositories.

Individual users also bring about change in the repositories, although those changes are primarily at the level of the dataset. Users detect errors in data, alerting repository staff to problems that need correcting. They request data formats for their own use, increasing the range of data types available at the repository. Finally, particularly at UMMZ, users are an important source of expertise driving the changing understanding of data. Given the vast number of specimens at UMMZ and the limitations on internal expertise corresponding to staff interests at a given time, the knowledge of zoologists using the collection is an important source for understanding the specimens.

6. CONCLUSIONS

An ongoing challenge for ICPSR, Open Context, and UMMZ is to balance the need for fixity in the datasets they offer with the fluidity of changing practices in their designated communities. As the contexts of data creation, sharing, and reuse change, these repositories vary the means by which they process datasets and make them available. In some cases, changes in the designated community can change the meaning of a dataset and its representation as well.

Repositories have some methods for ensuring fixity in the face of all this change. The professional practice of repository staff dictates the use of data standards like DDI and ArcheoML, which ensure some consistency over time. The repositories use a kind of controlled change to manage fluidity: ICPSR staff members check back with data producers as they process datasets, Open Context versions their datasets so that they can be retrieved in their former states, and UMMZ uses caution when responding to new species identifications in the physical arrangement of their collection, waiting until the designated community reaches a general consensus.

In this paper we identified a number of changes that repositories make in the process of working with data. At the level of data, they add value, correct errors, create consistency between datasets, and change the representation of datasets as their meaning changes. At the level of repository policy, they change collecting focus and data formats in response to changes within the designated community. While practices within designated communities undergo ongoing change, repositories respond by changing their own practices, but without a clear understanding of what future, as yet unseen, changes in the designated community may bring. This ongoing tension between fixity and fluidity guides much of the work of these organizations.

7. ACKNOWLEDGEMENTS

This research was made possible by a National Leadership Grant from the Institute for Museum and Library Services, LG-06-10-0140-10, “Dissemination Information Packages for Information Reuse”.

REFERENCES

[1] Allison, A., Currall, J., Moss, M. and Stuart, S. 2005. Digital identity matters. *Journal of the American Society for Information Science and Technology* 56(4), 364–372. DOI= 10.1002/asi.20112.

[2] Bolter, J. D. 1991. *Writing Space: The Computer, Hypertext, and the History of Writing*. Hillsdale, New Jersey: Lawrence Erlbaum.

[3] Brown, J.S., Duguid, P. 2000. *The Social Life of Information*. Boston: Harvard Business School.

[4] Coyne, M., Duce, D., Hopgood, B., Mallen, G., & Stapleton, M. 2007. *The Significant Properties of Vector Images*. Retrieved April 4, 2011 from http://www.jisc.ac.uk/media/documents/programmes/preservation/vector_images.pdf.

[5] Dappert, A. Farquhar, A. 2009. Significance is in the eye of the stakeholder. In *Proceedings of the 13th European conference on Research and advanced technology for digital libraries*. ECDL'09. Springer-Verlag Berlin, Heidelberg: 297-308.

[6] Duke, M, Day, M. Heery, R. Carr, L. A., Coles, S.J. 2005. Enhancing access to research data: The challenge of crystallography. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL ’05. ACM, New York, NY, USA, 46-55. DOI= 10.1145/1065385.1065397.

[7] Heslop, H., Davis, S., Wilson, A. 2002. *An Approach to the Preservation of Digital Records*. Retrieved September 4, 2011 from http://www.naa.gov.au/images/an-approach-green-paper_tcm2-888.pdf.

[8] Kansa, E. 2005. A community approach to data integration: Authorship and building meaningful links across diverse archaeological data sets. *Geosphere* 1:97-109. DOI= 10.1130/GES00013.1.

[9] Kelton, K., Fleischmann, K. R., & Wallace, W. A. 2008. Trust in digital information. *Journal of the American Society for Information Science and Technology* 59(3), 363-374. DOI= 10.1002/asi.20722.

[10] Levy, D. 1994. Fixed or fluid?: document stability and new media. In *Proceedings of the 1994 ACM European Conference on Hypermedia Technology*, ECHT ‘94. ACM, New York, NY, 24-31.

[11] National Archives (UK). 2002. *Defining the Characteristics for Authentic Records*. Retrieved September 6, 2011 from http://www.nationalarchives.gov.uk/documents/generic_reqs1.pdf.

[12] *Reference Model for an Open Archival Information System (OAIS).* 2002. CCSDS 650.0-B-1; Consultative Committee for Space Data Systems: Washington, DC.

[13] *Reference Model for an Open Archival Information System (OAIS).* 2009. Draft Recommended Standard. CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1 August 2009. Consultative Committee for Space Data Systems: Washington, DC.

[14] Vardigan M, Whiteman, C. 2007. ICPSR meets OAIS: applying the OAIS reference model to the social science archive context. *Archival Science* 7:73–87. DOI= 10.1007/s10502-006-9037-z.