Abstract: Migrating Research Data Collections (RE-07-18-0118)

"Migrating Research Data Collections" is a three-year (November 2018 – October 2021), $428,935 Early Career Development research project led by Dr. Andrea Thomer. In this project, the research team will investigate questions related to the migration of research data collections between different management and preservation platforms. The project will develop case studies of research data collection migration; a model of data collection migration patterns; best practices; open access course modules; and scholarly publications.

Broad Need. Implicit in the work of digital curation is the recognition that digital collections must last longer than the infrastructures on which they are stored. Migration of digital collections from one platform to another is therefore a fundamental aspect of curatorial work – yet, there is surprisingly little guidance for information professionals faced with this task. While metadata standards, interoperability guidelines, and careful selection of preservation-ready file formats certainly render *individual* digital objects more migration-ready, they don't necessarily support curators in the complex tasks (e.g. data modeling, schema matching, data transformation, provenance capture, etc) entailed in migrating an entire collection. Similarly, many best practices are focused on making individual datasets fit for a specific use or purpose, rather than a large collection of data, or collection of datasets, fit for ongoing curation. Work is needed to understand the unique data and database migration needs of information professionals, and to begin developing shared, open source, and community-driven tools and best practices which are appropriately tailored and scaled for their work.

This project answers the following questions:

- What patterns of use, access, and obsolescence drive the migration of research data collections, and how?
- What are the core curatorial processes that facilitate the migration of research data collections?
- What tools, guidelines and best practices are needed to better support this work in the future?

Outcomes include: 1) case studies of research data collection migration; 2) a model of migration patterns in digital collections – the rhythms of work and sociotechnical forces that trigger, delay, support, facilitate, and/or stymie migrations, as well as the specific curatorial processes entailed in different migration actions.; 3) drafts and ideas for best practices, guidelines and tools to support information professionals tasked with migrating research data collections; 4) open access course modules on collection migration; 5) scholarly publications.

Project design. We propose a unique combination of qualitative and participatory design methods: first, multi-site case study development, through which a model of collection migration patterns and processes will be iteratively developed, and secondly, a Community-Driven Design workshop in which study participants will be invited to contribute to the refinement of the model, and to identify priorities for guidelines and tool development. We ground our work in studies of Natural History Museums, because of their decades of experience in database development and migration. We then expand out to other LAMs. The proposed work takes place over three overlapping phases: *1) NHM case study development*; *2) Comparative case study development*; *3) Community engagement through participatory design.*

Diversity plan. Through all phases of work, this project will information professionals at a range of career stages, settings and scholarly backgrounds. The team will recruit a gender-balanced and racially diverse group of study participants, and a culturally and geographically diverse group of sites.

Broad impacts. There is an urgent need to improve upon the migration of research data collections, and the proposed work will have immediate impacts in LIS practice, education, and theory. This work will improve upon curatorial practices for LAM information professionals managing data collections. Additionally, it will build community by bringing together previously disparate groups and breaking down the historical "silos of the LAMs." Educational materials developed will improve upon curatorial practices and professional training.

Migrating Research Data Collections (RE-07-18-0118)

Overview

Implicit in the work of digital curation is the recognition that digital collections need to last longer than the infrastructures on which they are stored. Memory institutions such as museums, archives, libraries and research data repositories aim to create digital collections that last decades or longer – yet they must necessarily rely on hardware and that is designed to last just years at best. Data migration ("the process of moving data from one information system or storage medium to another to ensure continued access to the information as the system or medium becomes obsolete or degrades over time" ("migration," n.d.)) is therefore inevitable, and a fundamental aspect of curatorial work. However, there is surprisingly little guidance for information professionals faced with this task. While metadata standards, interoperability guidelines, and careful selection of preservation-ready file formats certainly render individual digital objects or files more portable from one system to another, these approaches don't necessarily support the migration of *collections* of data, in which the complex relationships *between* objects or records must be maintained. Similarly, many best practices are focused on making individual datasets fit for a specific use or purpose, rather than a large collection of data, or collection of datasets, fit for ongoing curation. Work is needed to understand the unique data and database migration needs of information professionals, and to develop shared, open source, and community-driven tools and best practices which are appropriately tailored and scaled for their work.

In this three-year, Early Career Development research project, Dr. Andrea Thomer and student researchers will investigate questions related to the migration of research data collections between different management and preservation platforms over the course of their lifespan. We ground our work in studies of Natural History Museums, thereby leveraging the NHM community's decades of experience in database development. We then move to other LAM sites. Research questions include:

1. What patterns of use, access, and obsolescence drive the migration of research data collections, and how?
2. What are the core curatorial processes that facilitate the migration of research data collections?
3. What tools, guidelines and best practices are needed to better prepare information professionals for this work?

In answering these questions, the project will develop 1) case studies of research data collection migration; 2) a model of migration patterns in digital collections; 3) drafts of best practices and ideas or prototypes of tools to support information professionals in their migration work; 4) open access course modules on collection migration; 5) scholarly publications. This work will make important contributions to the theory and practice of library and information science and related fields.

1. *Statement of Broad Need.*

Where library, archive and museum (LAM) collections catalogs were once primarily used as an access point to physical collections, more and more LAMs are realizing the importance of making their digital collections and metadata catalogs available as scholarly resources in their own right. For instance, the Cooper Hewitt Museum, Digital Public Library of America, and numerous other institutions are now making their collections data available through groundbreaking APIs (Chan, 2014; "DPLA API Codex," n.d.); museums like the Tate Modern have released their collection data and metadata, which has subsequently been studied and remixed in dozens of projects and publications (Tate, 2013/2018); and large consortia such as the Hathi Trust Research Center have been developing innovative tools to support the analysis of digital library holdings even when copyright restrictions prohibit direct access (Plale et al., 2013). These and many others have joined the call to consider "collections as data": the idea that digitized and born digital collections can be treated as "data rather than simple surrogates of physical objects or static representations of digital experience" (Padilla, 2018).

A similar "collections as data" paradigm has long shaped digital curation work in Natural History Museums (NHMs), though due to historical and domain differences in professional training and dissemination, these efforts have largely been isolated from other LIS and LAM endeavors. As early as the 1960s, NHMs began "computerizing" or "digitizing" (manually transcribing) their paper and card catalogs into databases, and by the 1970s began accessing and sharing records via information systems such as SELGEM and TAXIR, two information retrieval systems for taxonomic data[1] (Hudson, Dutton, Reynolds, & Walden, 1971; Mello, 1975; Sarasan, Neuner, & Association of Systematics Collections, 1983). Numerous other community-driven data and collections management platforms have been developed since then (for instance, Specify and Arctos, two NHM-specific collections management databases), as well as numerous other data sharing networks (for instance, MaNIS, the Mammal Networked Information System (Stein & Wieczorek, 2004); FishNet, a data sharing network for ichthyological collections (Vieglais, Wiley, Robins, & Peterson, 2000); and most recently and well-known, GBIF, the Global Biodiversity Information Facility (Robertson et al., 2014). It is estimated that the grand "dataset" of all NHM collections contains anywhere from 1-2 *billion* specimen records, only a small fraction of which are currently accessible through data sharing platforms (Ariño, 2010).

This deluge of research data collections has created an urgent need to reconsider best practices in digital collections management. Collections are typically intended to last for generations – yet *digital* collections must necessarily rely on hardware that only lasts years at best. Research in digital curation and preservation has made substantial strides in bridging the gap between the long lifespans of digital collections and the short lifespans of their infrastructures, as university libraries are increasingly embracing their central role in the management and preservation of "long tail" research data produced by faculty and students (Heidorn, 2008). For instance, models of curatorial and preservation processes have been developed and implemented as conceptual backbones to curation work (e.g. CCSDS, 2012; Higgins, 2008). Domain analysis, studies of disciplinary data practices, and research into scientific information work have contributed to our understanding of user needs and practices, and facilitated the development of robust, usable data stores and repositories (e.g. Borgman, 2010; Nielsen & Hjørland, 2014; Palmer, 2006; Palmer & Cragin, 2008; Weber, Baker, Thomer, Chao, & Palmer, 2013). Numerous metadata standards have been developed to facilitate the sharing, reuse and interoperability of data and metadata (Chen, Alderete, & Ball, 2016). Statistical modeling approaches aim to predict obsolescence rates in file formats (Duretec & Becker, 2017). Conceptual analytic approaches have shown how the formal definition of entities such as "dataset" and "document" impact our ability to manage, integrate and describe these objects in a digital environment (Furner, 2016; Renear & Dubin, 2003; Renear, Sacchi, & Wickett, 2010; Wickett, Sacchi, Dubin, & Renear, 2012), and defined the logical relationship between collections and their items (Wickett, in press; Wickett, Renear, & Furner, 2011).

As impactful as these contributions are, though, many have overlooked a critical component of digital collections management: the maintenance of the collections infrastructure itself. By and large, data curation best practices are designed to support the management of individual items. Recommendations for file format selection focus on the formats of individual data objects; standards such as Dublin Core are explicitly designed to describe one and only one file; and curatorial lifecycle models primarily serve to guide the curation of individual objects through a curatorial system. There is little guidance for the maintenance of the infrastructures on which research data collections are stored.

One of the most fundamental processes in data collection maintenance is *database migration:* "the process of moving data from one information system or storage medium to another to ensure continued access to the information as the system or medium becomes obsolete or degrades over time" ("migration," n.d.). Just as physical collections must be periodically reorganized and rehoused in response to shifting user needs, evolving curatorial priorities, space constraints, and material decay, so must digital objects be migrated from one database to another. However, where analog "data" migration is a well understood process, in a digital collection best practices guiding these processes are still nascent. Migration from one database system to

[1] Biological taxonomy, sometimes called systematics, is the discipline of biology concerned with the naming and classification of organisms.

another requires advanced knowledge of database systems, information modeling, and data cleaning. The underlying data models (the representation of information within a database and the relationships between different kinds of information) might need to be altered in a new system, or certain data fields may need to be reformatted or otherwise cleaned. Best practices in digital preservation do not necessarily guide curators in these complex tasks.

While there are certainly best practices aimed at supporting enterprise-level relational database migrations (e.g. Brodie & Stonebraker, 1995; Henry, Hoon, Hwang, Lee, & DeVore, 2005; Vassiliadis, 2009), and there has been substantial research in the field of computer science on database migration (work by Peter Buneman and collaborators is particularly relevant: Buneman, Chapman, & Cheney, 2006; Buneman, Cheney, Tan, & Vansummeren, 2008; Buneman, Müller, & Rusbridge, 2009) these approaches do not easily scale down to LAM or many scholarly contexts. Many enterprise-level approaches require a level of financial, technical and organizational support for database work that unfortunately isn't available in budget-strapped LAMs, where databases are often managed on an *ad hoc* basis by information professionals with a number of other work responsibilities. Additionally, LAM research data collections are more likely to be distributed over a number of idiosyncratically-structured legacy databases (which may or may not be relational), or "off-the-shelf" platforms over which data curators have little direct control.

Current best practices also do not necessarily assist information professionals in planning a database migration or selecting platforms that best fit their institutional settings and long-term priorities. Migration planning requires an understanding of the likely rate of obsolescence of hardware and software, as well as awareness of community trends and needs. Information professionals faced with database migrations need guidance navigating the complex sociotechnical factors that shape decisions about software and hardware purchases, as well as curatorial priorities. Research by scholars in science and technology studies and computer-supported cooperative work has begun exploring the role of databases in the scholarly workplace (e.g. (Bietz & Lee, 2009; Bowker, 2000; Hine, 2006; Manovich, 1999), but further work is needed to tease apart these forces in a way that provides information professionals with actionable guidance. As more and more memory institutions adopt "off-the-shelf" platforms (e.g. Fedora, DSpace, Islandora, KE EMu, etc.) over which they have little control, research is needed to understand the impacts these technologies have on the long-term sustainability of collections as well as the division of labor (and control) between information professionals and developers.

So long as we seek to store generation-scale collections on hardware not designed to last a decade (at best), we will need to migrate our digital collections. Given the rapid growth of "collections as data" initiatives, this work will only become more urgent. We posit that the migration of research data collections happens rarely enough (perhaps once every 5 years) that it has previously been often overlooked in best practices development as not being of pressing or day-to-day concern. However, at the point of migration, the need for evidence-based approaches becomes critical. We do note that while little formal research on the migration of data collections has been conducted, several case studies have been published by practitioners, and database migration is a frequent topic of conversation on community listservs. While informative, these case studies and email threads are too scattered to support theory building and the development of robust guidelines. Thus, there is a pressing need for empirically-based LAM-centric best practices in migrating research data collections.

Through the proposed project, we will conduct the research needed to ground these best practices. We propose to collect case studies of research data collection migration in a range of LAMs. We will identify common *migration patterns* in research data collections: the rhythms of work and sociotechnical forces that trigger, delay, support, facilitate, and/or stymie migrations, as well as the specific curatorial processes entailed in different migration actions. This research will act as a map the unique information ecologies (Nardi & O'Day, 1999) that contextualize LAM research data collections: the varied pieces of hardware and software used to manage our collections; the many digital and physical artifacts they contain and describe; the diverse uses and users of these platforms; the policies and regulations guiding and constraining use; and the relationships between these entities. Only once we have understood and developed theory around research data collection migration as it is currently conducted can we begin to build tools and best practices to support this work.

Pilot work: migrating NHM research data collections. The PI of the proposed project has recently completed pilot work in the context of natural history research collections. A paper describing this work has been accepted to the upcoming annual conference of the Association for Information Science and Technology (A. K. Thomer & Twidale, 2014; A. K. Thomer, Weber, & Twidale, 2018). NHMs were selected as a grounding case for this work. As early adopters of database technologies, they offer a unique longitudinal view on data curation and database management. NHMs are additionally informative for their use of relational databases for data management. Relational databases have been fixtures in many offices and research labs for decades (despite predictions of their impending obsolescence (e.g. Atzeni et al., 2013)), and they play an important, complex and often shifting role in information ecologies (Hine, 2006, p. 200).

The PI interviewed 12 collections managers (CMs), researchers and curators at a range of NHMs. Participants described 37 databases in total, with a median of 3 per department. Though we expected to find some instances of active database migration, we were surprised by how many of our participants were actively engaged in migration. At the time of our interviews, all but 1 collections database described by our participants was in the process of being migrated or was being prepared for a migration in the near future. An additional 2 CMs were in the midst of planning migrations of their *physical* collections. Thus, we find that database migration is not a single activity undertaken within a constrained time period, but instead, is an on-going (albeit often interrupted) aspect of their day-to-day work.

Participants drew on a range of techniques to manage their data and databases. We have distilled these into the following *migration patterns:*

- *Strategic denormalization.* Several participants deliberately denormalized databases to create "safe zones" for data entry, or otherwise keep similar data separated for access by different user groups. In these cases, CMs are essentially adapting the mechanics of relational databases to support security, privacy or usability needs.
- *Co-opting fields*. Users of databases with pre-set schemas and data entry interfaces sometimes co-opted fields for other-than-their-intended purpose. By doing so they are able to alter the database for their particular context without having to make changes at the schema level.
- *Reworking relationships.* In several cases, CMs had to transform the relationships between their records and their specimens, or between different tables in their data schemas. In our cases, this work was largely done "by hand" – through manual manipulation of columns and rows within tables. Though there are tools for schema migration work they are complex and aimed at users with substantial expertise in computer science. The commonality of this kind of work implies a need and market for tools to assist users in the complex work of data schema manipulation, and entity-relationship reworking.
- *Community consultation, and duplication with adaptation.* All of our participants spoke extensively with their fellow collections managers before beginning a database migration or selecting a new database system. Further, two of the legacy databases were found to have been copied from the same ancestral structure. NHM databases have been developed within communities even before the advent of systems like Arctos and Specify. Though this kind of memetic transfer of information models can lead to some problems – key details are lost through every transfer, as in a game of "telephone" – it's important to note and respect the role that community support and advocacy plays in selecting and maintaining research data systems.
- *Reliance on spreadsheet software.* Where others have noted the impact of the "psychological heritage of print" on users' wayfinding in databases (Kerr, 1990), here we note the psychological heritage of the spreadsheet for data entry and manipulation. Numerous participants used spreadsheet software as a place to

temporarily store data (though we note that some took a fairly long view of temporary – months to even years at a time) between migrations. Others use them as a *lingua franca* to ease collaboration between near and distant project partners.

This pilot work confirmed the centrality of data collections migration in collections management; indeed, it revealed that migration planning is an on-going concern for CMs, and that small steps toward facilitating migrations are embedded in their day-to-day work. Further, it confirmed that NHMs are an excellent grounding case in which to base further research. We suspect that migration patterns identified above will translate to other contexts, though further work is needed to validate this claim.

2. *Project Design*

The proposed project aims to scale up from pilot work to develop theory, and then action steps for, supporting the work of research data collection migration. We propose a unique combination of qualitative and participatory design methods: first, multi-site case study development, through which a model of collection migration patterns and processes will be iteratively developed; and secondly, a Community-Driven Design workshop in which study participants will be invited to contribute to the refinement of the model, and to identify priorities for guidelines and tool development. We will continue to use NHMs as our grounding case but then branch out to other LAMs. Throughout this project, we will develop educational materials to be shared with the broader LIS community and integrated into the PI's teaching. This research will make both theoretical and practical contributions to LIS. We ask the following research questions (RQ):

1. What patterns of use, access, and obsolescence drive the migration of research data collections, and how?
2. What are the core curatorial processes that facilitate the migration of research data collections?
3. What tools, guidelines and best practices are needed to better prepare and support information professionals in this work?

Methodological approach

Iterative, multi-site case study development (Addresses RQs 1 & 2). Case study methods are well-suited to research that aims to describe, explain or assess a phenomenon, particularly when that phenomenon cannot be easily disambiguated from its context (Yin, 2012). Given how deeply embedded – sometimes to the point of obfuscation – data migration processes are in the long-term work of collections management, case study methods are uniquely appropriate. Through semi-structured interviews with site stakeholders and content analysis of key documents and artifacts, the PI and student researchers will reconstruct the histories of collections databases, their migrations, and the sociotechnical arrangements around their maintenance over time. We will use methods of information modeling such as entity-relationship diagramming (in which the information classes of a database, as well as the relationships between classes, are modeled as a set of tables) and schema mapping (in which two versions of a schema are mapped to one another for comparison) to support our analysis and historical reconstruction (Figure 1). Interview protocols will draw on critical incident technique, in which participants are asked to describe a key point of transition in their custodianship of a system (Flanagan, 1954); this will help reveal the decision making factors, impacts and critical junctures in migration work.

Given budget and time considerations, only 3 sites will be visited in person (described further in *Specific activities).* Most cases will be developed remotely with interviews conducted by videoconference, as in the PI's successful pilot work. Atlas.ti or NVIVO will be used as a case study database; transcripts of interviews, documents and other evidence will be stored and coded by all team members to identify migration processes and other emergent themes, using a grounded theory approach (Strauss & Corbin, 1994).

Case study development will take place over two phases of work. Phase I will develop 3 initial cases of research data collections in or derived from NHMs, thereby building on already completed pilot work, capitalizing on the decades of experience NHM practitioners have in data collection migration, as well as leveraging the PI's extensive ties with this community. Phase I case study development will entail interviews with 5-8 information professionals at a site. Most interviews in Phase I will be face-to-face; this will help student researchers build confidence and expertise in interviewing techniques. Phase II will scale up from this work and develop 8-10 lighter-weight cases of data collection migration beyond NHMs (described further in *Site and participant selection).* Because these cases are intended to extend or refine our findings, they will be smaller in scope than Phase I; each case study will entail interviews with an estimated 3-5 stakeholders at a site.

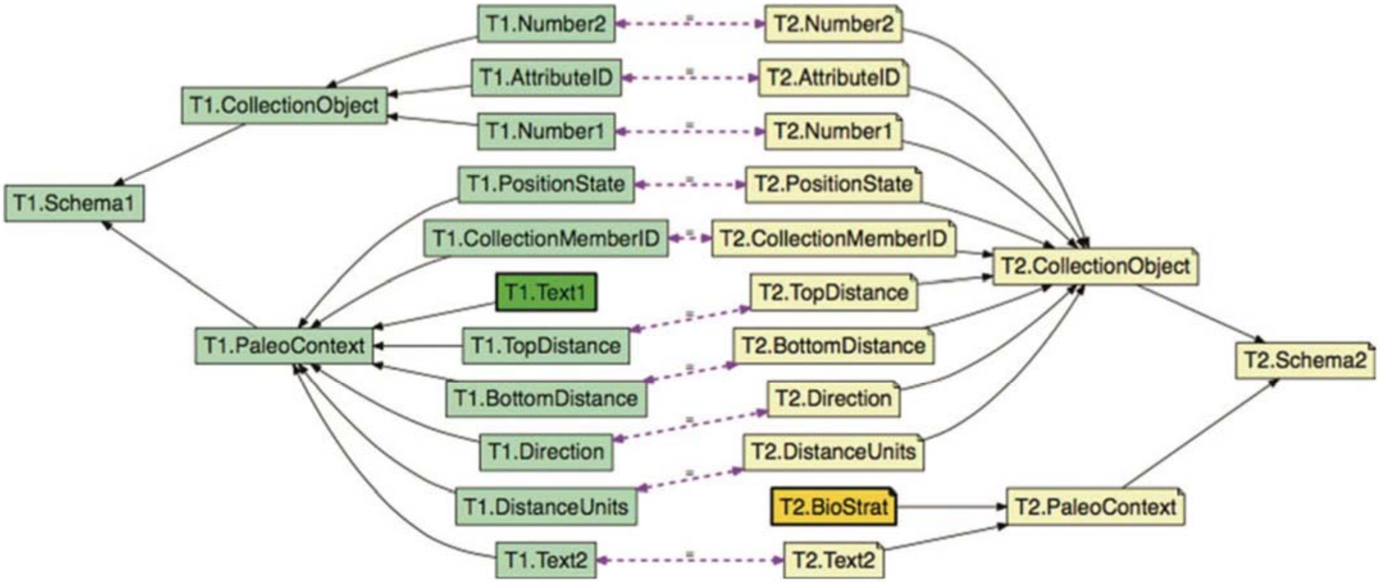

Figure 1. A schema map comparing two versions (version "t1" on the left and "t2" on the right) of a collections database over time. In this diagram, we see that one field from schema T1 (Text1) is absent from schema T2, and that T2 has an additional new field labeled "BioStrat." This and similar information models are used to visualize otherwise obscure changes in a database's structure over time. From Thomer, Cheng, Schneider, Twidale, & Ludäscher, 2017.

Intercase analysis; iterative identification of migration patterns (RQs 1 & 2). Concurrent with case study development, common migration patterns and processes will be identified via intercase analysis (Yin, 2012) and grounded theory methodology. As each case is completed, themes and patterns that emerge from coding will be compared to those identified in prior cases. A typology of the kinds of work that go into migrating research data collections will emerge. It is anticipated that this typology will be similar in structure to the patterns presented above from pilot work, and will include descriptions of the kinds of transformations that must be made to data structures and schemas (e.g. merging or splitting tables; merging or splitting fields; exporting and cleaning data); common obstacles faced and surmounted (e.g. lack of edit permissions; a need for new hardware); and key considerations that must be evaluated before making a decision (e.g. user needs vs. curatorial needs; the balance between off-the-shelf vs. home grown systems).

Community engagement through participatory design; development of guidelines and educational materials (Addresses RQ 3). One of the goals of this work is to contribute to the development of tools and best practices to improve data collection migration practices. A Community-Driven Design workshop in Phase 3 will help us connect our research findings to the day-to-day practices of those already in the field. 10 external (e.g. outside of the University of Michigan) and 5 local participants will be invited to participate in a 2-day workshop at the School of Information. Thomer will draw on methods of participatory design, as well as her past experience leading similar workshops (Thomer et al., 2014; Thomer, Twidale, Guo, & Yoder, 2016). Prior to the workshop, student researchers will develop drafts of best practices and ideas for tools, based in our prior findings. At the workshop, we will present these drafts and designs to participants. The research team will then guide participants through further brainstorming, paper prototyping, and drafting activities. Select outcomes will be further developed by student researchers for the remainder of the grant. It is not expected that this design work will result in fully functional, deployable software, but rather, will provide an opportunity for research-through-design (Zimmerman, Forlizzi, & Evenson, 2007). The workshop will additionally act as an important community-building and dissemination outlet in and of itself; preliminary work has shown that information professionals working on data migration often lack community support in this work, and the workshop will be an important first step in building awareness of the prevalence of these activities. By working to build community, we can give back to the professionals who will be taking time out of their busy days to participate in our study.

Throughout the entire project, The PI will incorporate project findings into teaching modules, which will be deployed in her own courses (e.g., SI 666: Information Organization and SI 667: Foundations of Digital Curation for 2018-2020; and similar future courses in LIS and Museum Informatics). Course modules will be shared at the project's end for potential adoption at other universities and for in-house training in the field.

Expected outcomes. In summary, we will develop the following outcomes, which contribute to both the theory and practice of LIS: 1) 11-18 case studies of research data collection migration, which will be summarized and published; 2) a model of migration patterns in digital collections; 3) drafts of best practices and guidelines and ideas or prototypes of tools to support information professionals tasked with migrating research data collections; 4) open access course modules on collection migration; 5) scholarly publications.

Specific activities and dissemination plan. The proposed work takes place over three overlapping phases:
Phase 1 (months 0-7): NHM case study development; development of initial migration models. In the first 7 months of the project, the PI will recruit student staff; meet with her advisory board; and obtain IRB approval for the study. The project team will complete 3 initial case studies of database maintenance and migration in NHMs; code and analyze the first 3 case studies alongside results from pilot work; develop a first draft of the migration framework; and disseminate preliminary findings at practitioner-oriented conferences such as RDAP, ALA and SAA to gain feedback and identify potential sites for Phase 2.

Phase 2: Comparative case study development beyond NHMs; refinement of migration model. From months 7-26 the project team will identify additional potential sites; recruit and complete 8-10 new cases; code and analyze cases as they are collected; and iteratively refine the migration model based on case study results. 2 cases will be selected for site visits; the remaining cases will be developed remotely via teleconference. Interim findings will be disseminated in LIS research communities such as ASIS&T and iConference.

Phase 3: Community engagement through participatory design; development of guidelines and educational materials development. From months 23-36 the project team will develop drafts of best practices and ideas for tools, based in our prior findings. We will also plan and host Community-Driven Design workshop. 10 external and 5 local participants will be invited to participate in a 2-day workshop at UMSI in Year 3. Following the workshop, select tools and guides will be refined or prototyped by the student team throughout the remainder of the year. We will additionally prepare teaching modules for dissemination; final reports for IMLS; and papers synthesizing research findings. Scholarly outputs will be published in outlets such as JASIST, the Journal of Documentation and the ACM Conference on Computer-Supported Cooperative Work.

Site and participant selection. Phase I will develop 3 cases of NHM research data collection migration; this builds on pilot work and lays a foundation for Phase II. The PI has strong relationships with key stakeholders at each organization and has received enthusiastic assurances of access and support in this work (See *SupportingDoc1.pdf*). The three sites include:

1) The Matthaei Botanical Gardens and Nichols Arboretum (MBGNA). The MBGNA is a "living collection" of plants distributed throughout four properties and over 700 acres of land in and around the University of

Michigan. The MBGNA's digital data collections consist of tens of thousands of items and records in at least 5 database systems. Data files include specimen records describing the type, locality, and provenance of each plant in the gardens and arboretum, as well as images, associated genetic data, and other data files. Collections staff are presently planning a migration of legacy records and genetic sequence data to a custom-built geodatabase.

2) The University of Michigan Natural History Museums (U-M NHMs). The U-M is home to several research museums, including the Kelsey Museum of Archaeology; the U-M Museum of Anthropology; the U-M Museum of Paleontology; and the U-M Museum of Zoology. Each of these museums manage substantial digital data collections and catalogs. Over the last several years, museum curators have sought to unify the museums' collections databases to better facilitate unified search, but their first attempt to migrate the collections to a proprietary system failed. However, the museums are once again working together to coordinate migrations to two different systems, with the goal of linking them via a unifying search interface through U-M library.
3) The Neotoma paleoecology research database (https://www.neotomadb.org). Neotoma brings together thousands of specimen records and published paleoecology observations into one system, thereby facilitating new kinds of integrative research using the aggregated data. The database itself is a complex system that was developed out of several other databases (notably, FaunMap and MioMap, two smaller paleoecology efforts), and therefore is an excellent exemplar of a data collection that has been migrated, maintained and adapted over time. Neotoma managers plan to migrate to a new data model in coming years, and case study development will focus on documenting these efforts as well as historical migrations.

Candidates for Phase II sites will be identified through consultation with the advisory board and via Phase I dissemination efforts; sites will be selected based for their potential to expand or confirm some aspect of the migration model and potential to contribute to a diverse representation of LAM research data collections. We will particularly focus on recruiting cases of data collection migration in academic libraries; archives; cultural heritage institutions; and institutional repositories. A potential site must adhere to the following criteria: 1) must identify as a LAM or memory institution; 2) must share data catalogs or collections for computational or programmatic analysis and use; 3) must have attempted to migrate local data stores within the last five years or be planning a migration within the next 3 years.

Participants for the Community-Driven Design workshop will be selected from study participants based on their engagement in case study development activities and interest/availability in participating in such a workshop. We will recruit a gender balanced and inclusive pool of workshop participants, representative of the breadth of collections in our study (see *Diversity plan).*

Human subjects and confidentiality concerns. Each study site will be quite unique, and the particular details of that uniqueness will be critical to contextualizing the cases. Consequently, we will not de-identify the names of the sites unless requested by study participants. This is in keeping with best practices of case study research (Yin, 2009). However, individuals will be assigned coded pseudonyms to protect their personal privacy. There will still be potential for re-identifiability with this approach, but it would provide them with a simple shield to prevent study results from appearing in search results. Participants will be informed of this potential identifiability through the consent process. The research team will additionally share drafts this dissemination materials with participants to ensure that they are comfortable with how they are being presented, and to give them an opportunity to ask that sections be more thoroughly anonymized.

Generalizability of study results and criteria for success. Case studies are sometimes critiqued as being too singular to be informative for other situations or contexts; however, our use of a multi-site design; an iterative, grounded analytical technique; and participatory refinement and consultation will mitigate this concern to the degree possible. The iterative design will help ensure that our findings represent general truths and common practices adaptable to other institutions, rather than isolated edge cases; and the Phase III workshop will provide the research team with a valuable opportunity to validate our findings with a relevant and invested community of practitioners.

Research Team

PI. Dr. Thomer is an assistant professor of digital curation at the University of Michigan School of Information, with interests in data curation, information organization, museum informatics, and computer-supported cooperative work. Prior to her graduate work in information science, she worked as an excavator and *de facto* databases administrator at the La Brea Tar Pits and Museum; her interdisciplinary background makes her uniquely qualified to complete this research. In addition to her over ten years of experience working in and with NHMs, she has developed workshops on digital humanities data curation, conducted research in earth and biodiversity informatics, and employed an array of qualitative, scientometric, and information modeling methods in her research.

Graduate student research assistant. A GSRA will be recruited in Year 1; it is anticipated that this role will be filled by a PhD student and that their work on this project would contribute to their doctoral training. The student will be responsible for collecting and analyzing case study data; coordinating and arranging site visits and interviews; developing a model of migration patterns and processes; contributing to papers and other research products; and assisting in project coordination. (See *SupoprtingDoc2.pdf*)

Hourly graduate assistants. In years 1 & 2, one student will assist in data collection and analysis, and will lead work in "data wrangling" and schema mapping activities necessary to evaluate our study participants' data and data structures. It is expected that these positions will be filled by master's students, and that this work will contribute to their professional development and competency in data curation. In Year 3, two students will be hired to assist with workshop development and to prototype data collection migration and maintenance tools that results from the workshops. It is expected that these positions will be filled by master's students with interest in both data curation and human-centered design, and that this work will contribute to their professional development as well. (See *SupoprtingDoc2.pdf*)

Advisory board. Four experts in digital curation and research collections have committed to serving on the project advisory board: Dr. Christoph Becker (University of Toronto), Thomas Padilla (University of Nevada, Las Vegas), Deborah Paul (iDigBio, Florida State University), and Dr. Karen Wickett (University of Texas, Austin). Each offers expertise in distinct areas of data curation. The PI will consult with the advisory board through annual telecons or feedback on site selection and methods, and for evaluation of research findings. The research team will additionally consult with individual advisors as appropriate over the course of the project. (see *Resumes.pdf*)

3. Diversity Plan.

Through all phases of work, this project will engage information professionals at a range of career stages, settings and scholarly backgrounds. Some of these participants are members of groups that have been previously overlooked by LIS research and training and domain communities, for instance, museum collections managers and curatorial assistants without LIS degrees. While their work is fundamental to supporting any research done with a collection, they are often not included as authors in scholarly publications, and thus, are sometimes unacknowledged in their research communities. Further, they are sometimes left out of technical and decision-making conversations about information technology, despite their experience in managing collections. By involving them as participant-collaborators, particularly in Phase III, we will make visible some of their invisible work and ensure that their voices are included in important conversations about best practices development and beyond.

The team will work to recruit a gender-balanced and diverse (in terms of race, sexual orientation, geography, institutional affiliation, urban/rural, and budgetary) group of study participants. The PI will work with her advisory board to ensure that historically under-resourced or overlooked collections from a range of regions are considered as research sites, as well as those that serve under-represented communities. All project team members will be asked to attend training sessions and read research on working with diverse groups (e.g. Steele, 2011) prior to conducting research or running workshop sessions.

4. *Broad Impact*

The migration of research data collections is an emerging area of importance, and the proposed work will have immediate impacts in LIS practice, education, and theory. Grounding our research in the extensive databasing efforts of NHMs ensures that we build on their insights and experiences; expanding the model through additional non-NHM case studies ensures that our best practices will support the migration of collections between a wide range of formats and platforms. Additionally, by basing our work in a broad range of memory institutions, this project will weave together previously disparate threads of research and practice and contribute to breaking down the historical “silos of the LAMs” (Zorich, Waibel, & Erway, 2008). Data collection migration is a common concern amongst these groups, and this project will help bring together each group’s unique strengths and contributions in service of a common problem.

Practically, the case studies developed through this work will be an important resource for LIS practitioners in and of themselves. The model of migration patterns will give practitioners, researchers, and developers a common language and framework for the discussion and support of future collection migration initiatives, studies, and tools. The best practices and prototypes that emerge from the Phase III workshop will be openly shared and therefore can support further development going forward. Finally, project findings will be incorporated into open source course modules, which can be used by other LIS educators and thereby build awareness of this important area of data curation and contribute to the training of future information professionals. By involving site stakeholders as partners in research, the Phase III workshop will lay groundwork for further partnerships.

Theoretically, the proposed work will make significant contributions to research on digital curation, information organization and representation, database design and implementation, and computer-supported cooperative work. Research data collections are important yet complex entities; our proposed work defining and articulating their patterns of migration and maintenance will be critical to their continued development and sustainability. LAMs are increasingly grappling with issues of sustainability and this work is an important contribution to research in that vein. This project will establish data collection migration as a new area of research in LIS.

Finally, the proposed work will make important contributions to the career development and training of the PI, the GSRAs involved in the project, and the PI’s students. This project builds on and extends Dr. Thomer’s long-standing work in data curation and natural history museum informatics and will support her development as a scholar and leader in this field.

		YEAR 1 - 2018-2019												YEAR 2.												YEAR 3											
		Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	**Nov**	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	**Nov**	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
Activity	Pre-award	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Recruit students	■										■												■														
Consult with advisory board	■							■							■									■					■								■
Phase I																																					
MBGNA data collection	■	■	■																																		
U-M NHMs data collection			■	■	■																																
Neotoma data collection				■	■	■																															
Coding and analysis; case report write up			■	■	■	■	■																														
First draft of migration model						■	■	■																													
Dissemination of pilot findings	■																																				
Phase II																																					
Identify candidate sites								■							■																						
Phase II site recruitment								■	■	■					■	■	■																				
Case study development									■	■	■	■	■	■	■	■	■	■	■	■	■	■	■														
Iterative analysis; case report write up										■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■										
Refinement of migration model													■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■					
Dissemination of interim research findings																				■	■	■	■	■	■												
Phase III																																					
Drafting of best practices & select tools																								■	■	■	■	■	■								
Workshop planning																											■	■	■								
Workshop participant recruitment																											■	■	■								
Community building and design workshop																													■								
Refine workshop outcomes																													■	■	■	■	■	■	■	■	■
Dissemination of final findings																																■	■	■	■	■	■
Final IMLS reporting																																			■	■	■
Education materials development																																					
Incorporate findings into courses		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Refine eductional materials for sharing																																	■	■	■	■	■

DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

□ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

Primary data will include transcripts from interviews, observational notes from site visits and workshops, copies of databases and data schemas, and other contextualizing documentation provided by study participants. The primary digital products from this work include scholarly manuscripts (such as peer reviewed journal articles, conference proceedings, book chapters, presentations, and reports); case studies of research data collection migration; drafts and designs of best practices, guidelines and tools to support future research data collection migration; and open access course modules based in this work, to be shared freely with the broader LIS community and integrated in the PI's teaching.

The investigators retain the copyright over the dataset and publications through the end of the project, but the final anonymized dataset and pre-prints of all scholarly manuscripts will be archived will be deposited with Deep Blue Data, the University of Michigan's institutional repository within 5 years; qualitative data may be archived with the Interuniversity Consortium for Political and Social Research (ICPSR) pending participant consent. I support reuse and remixing of data products and publications and will pursue the use of Creative

Common Attribution 4.0 license (CC- BY) when possible. I will also publish in open access venues when appropriate and available.

To support re-use, I will create clear metadata for all products, explaining data collection methods, attribution recommendations, and license details.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The University of Michigan will impose no condition on access or use of materials made available through the Deep Blue repository. Materials will be released with a CC-BY license when possible; given that much of this data will represent scholarly outputs, it's appropriate that we ask for attribution by re-users.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

This project poses no more than minimal risk to its participants. However, all reasonable steps will be taken to ensure participants' anonymity in research results. The names of participants in raw data and case studies will be anonymized of pseudonymized prior to data sharing or publication. The names of case study sites will be published if and only if study participants consent to their being shared. Risks of identifiability will be explained to all study participants through consent forms and through conversations with the project team prior to research activities.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

Primary digital products from this work include scholarly manuscripts (such as peer reviewed journal articles, conference proceedings, book chapters, presentations, and reports); case studies of research data collection migration; drafts and designs of best practices, guidelines and tools to support future research data collection migration; and open access course modules based in this work, to be shared freely with the broader LIS community and integrated in the PI's teaching.

File formats will include text files, audio files from interviews, spreadsheets, database files, metadata records in a range of formats (e.g. XML, JSON, XLS), and potentially computational scripts in languages like Python. Approximately 11-15 case studies will be collected; based on pilot work, it is estimated that each case will generate upwards of 25 data files (3-5 interview recordings (MP4) and transcripts (TXT); and a range of documents (DOC, PDF), images (JPG, PNG), and database files and records (MDB, SQL, TXT, XML, etc).

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

The research team will use computers and software provided by the University of Michigan School of Information.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

XML, RDF, XLS, JSON, XLS, CSV, MP4, TXT, DOC, PDF, JPG, PNG, MDB, SQL, and similar.

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

The PI will take primary responsibility for reviewing all project data prior to inclusion in publications, and prior to sharing. Graduate student assistant will assist in this work.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

Case study evidence will be managed and analyzed using qualitative research software such as Atlas.ti and NVIVO, as well as U-M Box.com cloud storage. After the award period, anonymized case reports will be deposited in the U-M's Deep Blue Data repository. Raw data may be deposited in ICPSR, pending participant consent.

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

The project team will work with data curators at Deep Blue Data and ICPSR to create appropriate metadata (likely in Dublin Core) for raw data and analyzed data products. Dublin Core is appropriate given the lack of standards specifically for case study data.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

Metadata during the project will be created and managed through NVIVO or Atlas.ti. Metadata after the project will be created with the assistance of data curators at Deep Blue Data and ICPSR, and using their tools.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

I will rely on the information retrieval and search functions of the repositories I deposit my data in. I will disseminate my findings and products through conferences and publications.

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

Part III. Projects Developing Software

Section not applicable to this proposal.

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

URL:

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

> This project will generate 8-11 case studies of research data collection migration. These will be developed through interviews, site visits, content analysis, and information modeling throughout the course of the three year project.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

> This data collection does require approval by the University of Michigan IRB; pilot work has already been approved; approval for the full study will be obtained in the months prior to the project's funding data.

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

The names of the individual participants will be anonymized or pseudonymized, but in keeping with best practices for developing case studies, institutional and departmental affiliations will not be changed. This increases the risk of individual identifiability for our participants. Participants will be made aware of this risk prior to their involvement of this work, and will be given the opportunity to opt out of research. Copies of papers and other dissemination materials will also be shared with participants prior to their publication to ensure they are comfortable with how they are being represented.

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

A codebook linking names to pseudonyms/codes will be maintained and stored separately from interview data. This will facilitate the linking between consent agreements and data.

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Case studies will be developed through interviews, site visits, content analysis, and information modeling throughout the course of the three year project.

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

A case study database will be maintained in NVIVO or Atlas.ti. Field notes, READMEs and metadata will be stored as plain text files along this database. Codebooks will be stored separately from this database to prevent accidental re-identification.

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Case studies will be summarized and published in a final report, and archived with the Deep Blue Data repository, ICPSR, or similar.

A.8 Identify where you will deposit the dataset(s):

Name of repository: Deep Blue Data

URL: https://deepblue.lib.umich.edu/data

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

The data management plan will be reviewed annually at the beginning of the academic year, to ensure that best practices are being adhered to.