University of Michigan, School of Information *Andrea K. Thomer, PhD*

Migrating Research Data Collections

"Migrating Research Data Collections" (MRDC) is a three-year, $499,890 Early Career Development research project led by Dr. Andrea Thomer at the University of Michigan. In this project, the PI and her students will investigate questions related to the migration of research data collections between different data management and preservation platforms over the course of their lifespan. It will also investigate the factors driving (and tensions arising from) the growing adoption of "off-the-shelf", sometimes proprietary, collections management databases. In answering these questions, the project will develop 1) **a model of migration patterns in digital collections**; 2) **best practices to better support the migration of research data collections**; 3) **prototypes of tools and techniques** to support migration processes – particularly focusing on migration audits, metadata creation, and data schema alignment; and finally 4) **open access course modules** based in this work, to be integrated in the PI's teaching and shared freely with the broader LIS community.

Statement of Broad Need. Implicit in the work of digital curation is the recognition that digital collections ought to 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.

Given the lack of best practices, information professionals have two choices: either develop *ad hoc* migration workflows and infrastructures or trust their collections to one of a growing number of standardized, often proprietary collections management databases. These options are less than ideal: *ad hoc* workflows are time consuming to develop and custom infrastructures are expensive to maintain; whereas migration to "off-the-shelf" systems (platformization) risks a loss of curatorial control and access. We posit that MRDC 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 point of migration, the need for evidence-based approaches becomes critical. Research is therefore needed to understand common processes and patterns in MRDC, and to develop best practices and tools for this work.

The PI of the proposed project has already begun conducting this research in the context of natural history research collections, because few information professionals know the challenges of MRDC so well as natural history museum (NHM) staff. As early adopters of computer-based collections management systems, NHMs have led the transformation of databases from simple utilities for search to sophisticated infrastructure for research. From 2014-2015, the PI interviewed collections managers at nine different institution, and found that every single participant was currently involved in a migration; that participants maintained a median of seven databases at a time; and that each database had been migrated a median of three times over its lifespan. Collection staff were frustrated at the lack of guidance in this work, but had developed a set of approaches common to their communities and likely applicable to others. Further work is needed to refine a model of collections migration, extend this model beyond the NHM, and develop best practices and training materials.

Project Design. This three-year project seeks to answer 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?
- How does "platformization" impact the accessibility and maintainability of research data collections? The use of customizable vs. standardized "off-the-shelf" systems? Open access vs. closed?
- What are the best practices for MRDC?

Specific activities are as follows:

Year 1: NHM case study development; development of initial migration models. The PI, a graduate student research assistant (GSRA), and 3-5 hourly student workers will develop three case studies of NHM database migration: 1) the Matthaei Botanical Gardens and Nichols Arboretum, as staff migrate legacy records and genetic sequence data to a custom-built geodatabase; 2) the U-M Natural History Museum, as staff migrate existing systems to an open source collections database; and 3) Neotoma (https://www.neotomadb.org), a paleoecology research database, as staff plan to migrate to a new data model in coming years. The PI has strong relationships with key stakeholders at all three of these organizations and has received enthusiastic assurances of access and support in this work. Case studies will be developed through site visits, semi-structured interviews with collections staff, content analysis of documents related to the migration process, and structured analysis of legacy and proposed data structures. The project team will draft an initial model of migration patterns and processes through inter-case analysis. Preliminary findings will be presented at practitioner-oriented conferences such as RDAP, RDA and ESIP to gain feedback and recruit participants for work in Year 2.

Year 2: Comparative case study development beyond NHMs; refinement of migration model. The project team will develop an additional 10-15 lighter-weight case studies of non-NHM research data collection migrations (e.g. scholarly collections in the digital humanities, social science data stores, and more). Year 2 site visits will be limited to just 3 exemplar cases; remaining cases will be primarily developed through semi-structured interviews (~2-3 per case). These cases will be used to iteratively test, refine and extend our initial model of migration patterns and processes. Participants will be recruited via professional networks, listservs, and social media. Results will be presented at outlets such as ASIS&T and iConference.

Year 3: Development of best practices & curriculum modules; tool prototyping. The project team will refine our migration model into a suite of best practices. We will prototype tools to support migration work. Specific tools will depend on student skills and project findings, but will likely range from scripts to facilitate tasks such as schema matching or data cleaning, to checklists to assess the migration-readiness of potential platforms. Project findings will be incorporated into open access teaching modules within the PIs curriculum, and shared for use by others in digital curation education. Results will be published in outlets such as JASIST and CSCW, with open access licenses where possible.

Diversity Plan. Through all phases of work, this project will engage students as well as information professionals at a range of career stages. The PI will leverage the University of Michigan School of Information's robust history of supporting diversity, equity and inclusion initiatives to recruit students from under-represented backgrounds in the information professions to this project.

Broad Impact. The proposed work will contribute to research on digital curation theory and practice, platform and infrastructure studies, computer-supported cooperative work, and knowledge organization. More practically: 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. This project will develop crucial insights and training for the work of MRDC. Rooting our models in the largely successful efforts of NHMs ensures that we build on their insights; expanding the model through additional non-NHM case studies ensures that our best practices will support the migration of collections between platforms as wide-ranging as Omeka to KE-EMu to MySQL.

Personnel and Budget Summary. The estimated budget is $499,890. IMLS Direct Costs include $45,644 Salaries/Benefits, $30,214 Non-Student Travel, and $272,139 Student Support Costs. (IMLS Direct Costs $347,997 + $151,893 IDC @ 56% = $499,890). There is no cost share requirement for this research proposal.

PI. Dr. Thomer is an assistant professor of digital curation at the University of Michigan School of Information, with interests in information organization, museum informatics, and computer-supported cooperative work. Her background is highly interdisciplinary; in addition to her over ten years of experience working in and with NHMs, she has conducted research and developed workshops on digital humanities data curation, and employs an array of qualitative, scientometric, and information modeling methods.