IJDC | *Conference Pre-print*

# Three Approaches to Documenting Database Migrations

Andrea K. Thomer
School of Information,
University of Michigan

Alexandria Rayburn
School of Information,
University of Michigan

Allison R. B . Tyler
School of Information,
University of Michigan

**Abstract**

Database migration is a crucial aspect of digital collections management, yet there are few best practices to guide practitioners in this work. There is also limited research on the patterns of use and processes motivating database migrations. In the "Migrating Research Data Collections" project, we are developing these best practices through a multi-case study of database and digital collections migration. We find that a first and fundamental problem faced by collection staff is a sheer lack of documentation about past database migrations. We contribute a discussion of ways information professionals can reconstruct missing documentation, and some three approaches that others might take for documenting migrations going forward.

*Submitted* 15 December 2020 ~ Accepted 19 February 2020

Correspondence should be addressed to Andrea Thomer, 105 S. State St., Ann Arbor, MI 48109. Email: athomer@umich.edu

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/



# Migrating Databases and Digital Collections

A fundamental aspect of digital collections management 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.). But while analog/physical "data" migration is well understood and theorized within LIS, digital collection migration is less well supported. In the "Migrating Research Data Collections" project (IMLS Grant RE-07-18-0118-18), we are developing these best practices through a multi-case study of database and digital collections migration. We are starting by developing case studies of database migration in natural history museums (NHMs), which are often overlooked in LIS research despite being early adopters of database technology and early contributors to scholarship in data curation (Palmer, Weber, Renear and Muñoz, 2013; Thomer, Weber and Twidale, 2018).

Here we present emergent finding from the first phase of this project: multi-site case studies of NHM collections migration at the Matthaei Botanical Gardens and Nichols Arboretum (MBGNA), the University of Michigan Natural History Museums (U-M NHMs), and the Neotoma Paleoecology database. Each of these organizations hosts long-lived digital data collections – well over four decades – and has consequently had to repeatedly migrate their data collections over the years. Though analysis is still on-going, our early work on this project illustrates some of the practical issues that information professionals face while planning and completing a data infrastructure migrations – as well as some more meta-level issues faced by data curation researchers seeking to study these changes over time. We find that one of the most common and fundamental problems information professionals face is a lack of documentation about past migrations, which can subsequently hinder current and future migrations. We contribute a discussion of ways information professionals can reconstruct missing documentation, and some guidelines for creating migration documentation going forward.

# Phase I: Case Studies of Research Data Migration at Natural History Museums

In this first phase of work, we are developing case studies of database migration in natural history museums (NHMs). Each case study is being developed through semi-structured interviews (45 to 75 minutes each) with curatorial and collection staff at each site; close analysis and comparison of different versions of legacy databases; and review of papers, memos, emails, and other documentation related to database migration. Evidence is triangulated to develop explanations of how and why migrations are necessary, and to identify patterns motivating migrations, following a multi-case study design (Yin et al., 2017). Cases 1 and 2 are complete; development of Case 3 will be complete by August 2019. Short summaries of each case follow, to be expanded on in the full paper and oral presentation.

## Case 1: The University of Michigan Matthaei Botanical Gardens and Nichols Arboretum

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 collections and catalogs date back to 1910, and their digital collections databases data back to the 1980s. The MBGNA's digital data collections consist of tens of thousands of items and records in several different 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. Current staff have a general idea of the digital collections' provenance: in the 1980s card catalogs were first transcribed into TAXIR; migrated to BG-Base in the 1990s; then to Microsoft Access in 2003; and updated to the most recent version in 2012. Since our initial interviews with MGBNA staff in 2018, the Access database has been migrated to an ArcGIS GeoDatabase. The migration to ArcGIS was unexpectedly challenging, though, because despite a general awareness of the database's history, there was little detailed documentation defining fields and relationships. Current staff had to essentially reverse engineer the database before migrating it into ArcGIS. In addition to the collections catalog data, MBGNA staff have also maintained separate data stores for specific gardens or individual field projects, which similarly require reverse engineering to migrate databases were created to, "suit [individual researchers'] own needs. And so, unfortunately as those staff members have left, we haven't always known exactly how or why those files were created" (Participant MBGNA-03). Complicating the lack of documentation has been historical siloization or territoriality over data: one participated described them as being managed, "as jewels of individual dragons, in terms of, 'This is my information, not yours'" (Participant MBGNA-01).

## Case 2: The University of Michigan Natural History Museums

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-M Museum of Zoology. Each of these museums manage substantial digital data collection sand 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 centralized databases. The biological and geological collections are being migrated to Specify, an NHM-specific collections management system. The anthropological and archaeological collections are being migrated to Collective Access, a customizable collections management system for cultural heritage systems. As at the MBGNA, U-M NHM collection staff similarly have been faced with the challenge of reconstructing long and often forgotten database histories to facilitate migration to new formats. Typically, the NHMs collection catalogs were also entered into digital databases in the 1980s but the specific details and those migrations have been lost to time – particularly for collections that have seen more staff turnover or less reliable funding than others. The "traces" left behind by legacy databases continue to impact current migration efforts – for instance, fields with unclear definitions, or that were split across multiple tables for unclear reasons.

## Case 3: The Neotoma Paleoecology Research Database

The Neotoma database brings together thousands of specimen records and paleoecology observations into one system, thereby aggregating data to facilitate new integrative research. Neotoma runs on a complex relational database that incorporates several older databases, which Neotoma managers are migrating to a new data model in coming years. Case study developments on-going for this case. Early work has shown, though, that while database managers here have a much clearer understanding of their systems' histories than in the prior two cases, they nevertheless lack some detailed documentation about field definitions, integrity constraints, and relationships.

# Discussion: Retrospectively Documenting Database Migrations

Our work makes several contributions for the digital curation community.

Though our development of these cases is still on-going, one commonality has immediately emerged: current collection staff are often tasked with managing data collections for which they do not know the precise origins or histories. This is in keeping with findings from plotwork for this project; in exploratory interviews with NHM collection managers, we found that they typically began their positions by essentially reverse engineering the databases they were charged with maintaining (Thomer et al., 2018). Without a clear understanding of current and legacy structures, information professionals run the risk of introducing new errors into a database during a migration. The collections staff we spoke with generally felt that their jobs would have been easier with more documentation of legacy data structures – but it is unclear just what form that documentation should take. Traditional methods of documenting database structure include Entity Relationship (ER) or UML diagrams: clear illustrations of the information classes, and the relationships between those classes, within a database. However, these diagrams don't show change over time. In developing each of the cases presented above, we developed three approaches to showing change over time that may be useful to both researchers and practitioners:

- Database Readmes. The narrative reports generated for each case represent a form of documentation in and of themselves, almost similar to a very extensive README file. The strength of this approach is its relative ease of creation: information professionals already have the skills needed to create qualitative, narrative histories of their information systems. However, narrative documentation runs the risk of being insufficiently precise.

- Versioned entity relationship diagrams. As noted above, ER diagrams are a familiar tool in database design. Creating versioned ER or UML diagrams is a clear way of documenting changes over time. However, ER diagrams require specialized training to create and can be challenging to compare, especially when databases have undergone extensive restructuring.

- Sankey diagrams. Sankey diagrams can be used to show the "flow" between information or resources. We have begun using them to show the evolution of data systems over time, and to visualize the flow between data stores. This is a less conventional approach to showing relationships between different versions of databases but may have promise in providing a big picture view of systems over time.

All of these diagramming methods are capable of acting as lasting documentation for a system. Additionally, they all have the potential to act as the "presentation view" that Jagadish has argued is necessary to give modern database users a clear understanding of the data models behind their systems (Jagadish et al., n.d.). In future work, we will be further refining these diagramming approaches, and working directly with study participants to test their efficacy as practical ways of documenting database change over time. We also hope to draw on work using logic-based schema alignment to visualize changes in the relationships between tables and fields between database versions (Thomer, Cheng, Schneider, Twidale and Ludäscher, 2017; Franz et al., 2015); this may be particularly powerful when paired with ER diagrams.

## Acknowledgements

This work was funded by IMLS Grant RE-07-18-0118-18. Thanks to our study participants frothier time and contributions.

## References

Franz, N. M., Chen, M., Yu, S., Kianmajd, P., Bowers, S. & Ludäscher, B. (2015). Reasoning over taxonomic change: Exploring alignments for the perelleschus usecase. *PLOSONE, 10*(2), e0118247. doi:10.1371/journal.pone.0118247

Jagadish, H. V., Chapman, A., Elkiss, A., Jayapandian, M., Li, Y., Nandi, A. & Yu, C. (n.d.). Making database systems usable. In SIGMOD '07: Proceedings of the 2007 ACM SIGMOD International Conference on the Management of Data (p. 13-24). ACM Press. doi:10.1145/1247480.1247483

migration. (n.d.). Society of American Archivists. Retrieved from https://www2.archivists.org/glossary/terms/m/migration

Palmer, C., Weber, N. M., Renear, A. & Muñoz, T. (2013). Foundations of data curation: The pedagogy and practice of "purposeful work" with research data. Archival Science. Retrieved from https://www.ideals.illinois.edu/handle/2142/78099

Thomer, A. K., Cheng, Y.-Y., Schneider, J., Twidale, M. & Ludäscher, B. (2017). Logic-based schema alignment for natural history museum databases. Knowledge Organization,44, 545–558. doi:10.5771/0943-7444-2017-7-545

Thomer, A. K., Weber, N. M. & Twidale, M. B. (2018). Supporting the long-term curation and migration of natural history museum collections databases. Proceedings of the annual meeting of the Association for Information Science and Technology, 55(1), 504–513. doi:10.1002/pra2.2018.14505501055

Yin, R. K. (2017). Case study research and applications: Design and methods. Sage publications.