Maintaining Repositories, Databases, and Digital Collections in Memory Institutions: An Integrative Review

Thomer, Andrea K. University of Michigan School of Information, USA | athomer@umich.edu
Starks, Joseph R. University of Michigan School of Information, USA | starksjr@umich.edu
Rayburn, Alexandria University of Michigan School of Information, USA | arayburn@umich.edu
Lenard, Michael C. University of Michigan School of Information, USA | mclenard@umich.edu

ABSTRACT
Database maintenance and migration are critical but under-supported activities in libraries, archives, museums (LAMs), and other scholarly spaces. Existing guidelines for digital curation rarely account for the maintenance needed to keep digital curation infrastructures functioning over time. Though many case studies have been published describing individual instances of migration, there has been little generalizable research done in this area. Thus, it is challenging to understand overall trends or best practices in this space. We bridge this gap by conducting an integrative literature review of papers describing database migrations and maintenance in LAMs and other scholarly contexts. By qualitatively coding 75 articles from 58 publication venues, we identify common motivations for database migrations and maintenance actions. We find that databases are migrated to support changing user needs as well as to ward off technological obsolescence; we also find that common challenges include schema crosswalking and a need for data cleaning. Practitioners describe community collaboration as key in surmounting these challenges. Through this integrative review, we build a base for further best practices development and identify a need to better model database curation as part of the digital curation lifecycle.

KEYWORDS
Data curation; database migration; sustainability; knowledge infrastructures; maintenance

INTRODUCTION
Implicit in the work of data curation is the work of data*base* curation: the on-going management and maintenance of the infrastructures needed to store, clean, and access digital objects in the long-term. Like all infrastructures, these systems have a tendency to fade into invisibility; we don't notice them until they break—or until they need to be updated, migrated or altered to fit changing user needs (Bowker & Star, 2000). Further encouraging invisibility, these infrastructures tend to change on longer, slower timelines than we are often accustomed to planning for in institutions and as data curators (Ribes & Finholt, 2009). Database maintenance thus tends to be omitted in discussion of digital curation.

The research on database migration that *does* exist is scattered throughout the academic and professional literature. For instance, there are a number of library, archive and museum (LAM) practitioner-authored papers describing "lessons learned" from individual system migrations, upgrades or other database renovations. Some research on digital curation includes limited discussion of infrastructure maintenance, often in the context of sustainability over time. And finally, there is some relevant material in the extensive computer science literature on database maintenance (e.g. Brodie & Stonebraker, 1995; Buneman et al., 2008; Jagadish et al., 2007) though this is largely dominated by highly technical discussions of database structure, indexing methods, and other issues more germane to the construction of systems rather than the maintenance of collections. Work is needed to synthesize this literature.

As part of a larger project in which we are studying database maintenance, here we systematically review this prior literature to build a base for further best practices development, and to better understand the state of the art of database migration. Our goal is to synthesize prior work—particularly case studies—and identify gaps and build a foundation for further best practices development and empirical research. Our research questions are:

- What are the primary themes or topical coverage of existing research on maintenance and migration of research databases and digital collections?
    - In what venues and communities is the work published, and by whom?
- What recommendations and best practices already exist to support the maintenance and migration of research databases and digital collections?

Our paper proceeds as follows: we first further describe our motivations in synthesizing research on database curation; we then describe the methods used for our integrative review of relevant literature; and we then describe

85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

key trends and "lessons learned." We conclude with a discussion of critical areas for future research in this area, notably a need to better understand users' changing needs of data infrastructure (and not just datasets) over time, and a need to build better tools to navigate the rigidity of data architectures over time.

MOTIVATION

Over the last three years, we have been studying the work of museum collection managers, librarians, archivists, and scientists as they maintain and migrate the systems used to manage digital objects and data (IMLS grant RE-07-18-0118-18; see Thomer et al., 2018; Thomer & Wickett 2020; Rayburn & Thomer, 2022). We are primarily interested in long-lived databases that are used within a memory institution as lasting infrastructure for some sort of collection—such as a museum catalog, a collection of scientific data, or the scholarly products deposited in an institutional repository. For simplicity's sake, we refer to this array of infrastructures simply as *databases*, but these systems go by many names and take many forms. Some (particularly in scientific data contexts and in natural history collections) rely on traditional relational database technology. Others (particularly in cultural heritage institutions) use complex digital asset or content management systems, which may include relational structures at some level, but also a broad array of other data structures and interfaces to store and provide access to digital objects. Some are used to manage detailed metadata cataloging physical objects such as specimens or books; others are used to manage digital objects such as digitized texts or digital photographs.

We group this broad range of collections and technologies together to accurately reflect the breadth of database-related work in memory institutions. We refer to this work as *database curation*, which includes both database maintenance and migration. By *maintenance* we mean making routine changes or updates to an infrastructure in response to some other pressure, such as changing user needs or shifting norms within a field; by *migration* we mean the movement of digital objects from one infrastructure to another. Maintenance work is often underfunded compared to the development of new projects or technologies (Jackson 2014; Russel & Vinsel, 2018), and we have frequently heard from the *de facto* database curators at LAMs that there is little support for this critical aspect of infrastructure sustainability—either institutionally or professionally. Thus, a main goal of our broader project is to develop frameworks to better illustrate and guide database curation.

A review of common digital curation lifecycle models and frameworks quickly confirms the lack of consideration for database curation. The DataOne data lifecycle (Cook et al., 2012), United States Geological Survey (USGS) data stewardship model (United States Geological Survey, 2014) and Digital Curation Center curation lifecycle model (Higgins, 2008) all center on the curation of individual data objects, but do not describe the maintenance or migration of underlying infrastructures. The IDCC model does include "migration" as one "occasional" curatorial action, but this again is migration of an individual dataset, not an entire system or collection. The Open Archival Information System framework (OAIS) (Consultative Committee for Space Data Systems, 2012) provides a more comprehensive view of a preservation system, but essentially black boxes the issue of infrastructure maintenance. This widely adopted, ISO-ratified framework for building an archival information system outlines the concepts and processes needed to ensure "long-term" preservation of archival materials ("long-term" is defined as a period of time "long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community."). Longevity of an OAIS is ensured primarily via the creation and maintenance of discrete "information packages" containing data and metadata; the OAIS framework describes multiple processes meant to ensure their continuity over time. However, OAIS does not consider the work needed to maintain or migrate aspects of the archival infrastructure itself—e.g., database software, servers, and so on. Thus, there is a notable gap in recommendations for practitioners faced with managing a data system—and not just datasets—over time.

Our goal in this paper is to surface the vital work of database curation by synthesizing the research that *does* exist on database migration and maintenance. Many LAM practitioners have published case studies describing individual database curation projects, and a few information science and disciplinary researchers have conducted more systematic research on this work. We aim to synthesize these findings, and thereby build a base for further guideline development.

METHOD

We conducted this review by developing and testing search queries to be used across several databases of literature, the results of which comprised the corpus for review. The corpus was practically screened by removing duplicates and non-English literature, performing a title and abstract review, and a full-text review. The remaining corpus was coded in NVivo using an inductive process, with second rounds of coding performed to discern motivations and lessons learned. We describe each step in this process further below.

Search and selection

We based our search and selection method on the guidelines described by Okoli and Schabram (2010). Working with a liaison librarian, we identified candidate databases to search for relevant LIS and CS literature. We tested our

sources and queries through iterative pilot testing, recording the total number results from a particular query applied to a particular database in a spreadsheet (Wolfswinkel et al., 2013). Effective query construction required an understanding of databases' index terms, thesaurus, and Boolean operators. Searches that produced excessively large numbers of results (e.g. greater than 500) in most databases were made more specific via the addition of related terms or disambiguating search phrases. For example, [database AND migrat*] could be improved by adding [AND "research data" NOT cities]. Furthermore, we eliminated sources that yielded few or irrelevant results regardless of the query. We judged the first 20 results for relevance at the title and abstract level, using the same criteria we would later use for practical screening, and recorded the number of relevant results. By comparing results of each search, we were confident that the sources chosen and queries used would provide appropriate literature and capture a broad corpus of relevant work.

Our final queries consisted of three main parts. The first part was either [migrat*] or [sustainable OR sustainability OR maint* OR curat*]. The second and third parts, which were the same for both queries, specified the types of systems we were interested in, such as [database] or [DAMS] or ["data repository"], and the contents of the system, including ["research data"] and ["digital collections"]. We altered the queries' syntax slightly for a few databases to ensure that they were applied in a logically consistent manner across all sources searched. When saving results for inclusion in the corpus, an upper limit of 60 items per search was chosen based on an observed sharp decline in relevance of results past the first 60 during pilot searching. After searching fourteen databases (ACM Digital Library; IEEE Xplore; Computer Database; Applied Science and Technology Abstracts; IET Inspec; Library and Information Science Abstracts; Library, Information Science and Technology Abstracts; Library and Information Science Source; Library Literature and Information Science Index; Library Literature and Information Science Full Text; Scopus; Web of Science; Google Scholar; and the Directory of Open Access Journals), we had gathered 841 items. The comprehensiveness of the search was corroborated by the high number of duplicate papers. The results were exported into Zotero where duplicates, work published prior to 2000, and non-English literature were removed.

We note a few limitations of our approach. First, we did not search databases that include books, leaving out database design textbooks and other volumes that likely describe some element of database maintenance, for instance, Brodie and Stonebraker's well known volume, *Migrating Legacy Systems* (Brodie & Stonebraker, 1995). Second, we additionally found that our query did not retrieve some computer science literature already known to us as related to database migration, such as Buneman et al.'s work describing curated databases (2008). We suspect this reflects the search limitations of the libraries we queried; some were restricted to abstract search rather than full text. Finally, we acknowledge that the focus on English language papers is a limitation of our work and leaves out a breadth of literature on the subject written in other languages. We look forward to future research analyzing works in other languages.

Review and analysis

Three members of the study team reviewed papers for relevance in our review. We first screened titles and abstracts for relevance, and then conducted full-text review of relevant papers (Okoli & Schabram, 2010; Wolfswinkel et al., 2013). Of the 841 papers retrieved by our query, 651 were unique documents written in English. We narrowed the corpus to 184 relevant items through title and abstract review, screening for items in journals, conferences, technical reports, white papers, and working papers about database migration, curation, and maintenance, digital collection migration and sustainability, and research database migration and sustainability. After a full-text review, screening for items that specifically discuss database system migration or maintenance, we were left with 75 items for coding and in-depth analysis.

We established inter-rater reliability at the title and abstract level via a random sample of 20 pieces of literature, rating articles for inclusion (yes/no) and discussing disagreement (Pati & Lorusso, 2018; Wolfswinkel et al., 2013). Standards for agreement in inter-coder and inter-rater reliability in qualitative research vary. McDonald et al. write that when coding and reviewing is the "process not the product" (McDonald et al., 2019), interrater reliability (IRR) is less necessary, provided significant and regular discussion takes place between the reviewers and the primary investigator. Landis and Koch, however, regard a kappa statistic of .61-.8 as "substantial" and .81 or greater as "almost perfect" agreement (Landis & Koch, 1977). We decided that 80% agreement, in combination with regular discussion, was sufficient. We proceeded similarly for a full-text review. We discussed disagreements about inclusion at the full-text level as they arose and the principal investigator served as the deciding vote.

We used NVivo to review and code the 75 relevant papers. We first categorized papers according to their approach (case study, other research project, position paper), setting (library, museum, research lab), and publication domain (library information science, computer science, domain publication, other). As coding progressed, we discussed observations as a team. We developed an initial high-level codebook based on our research questions, including: the authors' motivations for migration or tool development; any outcomes of the project; descriptions of specific

migration or maintenance processes and steps; lessons learned; and prescriptive recommendations for others engaging in similar projects. We found the *lessons learned*, *prescriptive recommendations*, and *motivations* sections to be particularly rich, and performed a second round of coding on these categories to identify common themes. This second round of coding helped us discern shared reasons for migrating and similar recommendations across the corpus, which are discussed in depth below.

RESULTS

Below we summarize the results of our literature review and synthesis. We cite papers from our review throughout the results, but a table including the full bibliography as well as classifications and coding results for each paper is available at https://doi.org/10.7302/5hr4-j779.

The corpus

Of the 75 relevant papers retrieved, the majority (n=57) are case studies describing data repository development, maintenance or migration at individual institutions. For instance, Knight-Davis et al. describe their experience of digitizing 80,000 herbarium specimens and migrating the collections metadata from Microsoft Access to Symbiota (Knight-Davis et al., 2015). Allen describes a case study of migrating multiple digital collections at the University of Arkansas to a single open-source institutional repository (Allen, 2017). Of the case studies, 13 describe the migration of a system as their main focus; 6 describe the maintenance of a system or its data in preparation for migration; and the remainder discuss other aspects of sustainability or infrastructure maintenance as part of a broader discussion of digital curation projects (e.g. Gentry et al. who discuss migration as a background concern in a broad digital and physical curation project (2021), or Knight-Davis et al. who discuss metadata migration in support of a digitization project (2015).

17 papers use other research methods (e.g. surveys, content analysis) to study database migration or maintenance more holistically (though we note that the majority of these broader research papers focus on general issues of infrastructure sustainability rather than maintenance and migration specifically). For example, Imker (2020) reveals major gaps in funding and maintenance in long-lived molecular biology databases by reviewing 67 such databases. She finds academic institutions bear the most responsibility for operational support after their initial funding through government channels. Imker concludes that mechanisms for redistributing these burdens would ease resource strain and improve sustainability. Rieh et al. conduct an interview study to better understand the perspectives and experiences of repository staff during institutional repository (IR) planning and implementation; they find that staff view IRs as more than the sum of their parts, and that designing systems around user needs is key to sustainability. In a paper from the computer sciences, Schuler and Kesselman propose a "schema evolution framework" for updating the schema and instance data of databases. The goal is to raise the level of abstraction to make complex database evolution tasks more accessible to scientific researchers. Finally, Kansa et al. (2005) describe efforts to make archaeological databases more interoperable; we classified this paper as research rather than a case study because they discuss obstacles to sustainability and interoperability more generally, and not several broader challenges related to schema migration that will be commonly faced by those working in relational systems.

Lastly, two papers retrieved are opinion or position papers, in which the authors advocate for some sort of policy, such a community platform and standardization in earth observation and environmental research data (Gries et al., 2018) or the long-term preservation of libraries digital collections (Breeding, 2002).

The majority of papers were written by LIS practitioners (such as database managers, librarians, repository staff, software developers, or curators) and published in LIS journals (such as the Journal of Librarianship and Scholarly Communication or the Journal of Library Metadata) (n= 43; Figure 1, left). An additional five (6%) were published in a LIS/ Archive and Record Management (ARM) conference proceedings. Fifteen (20%) were published in a computer science (CS) or data science venue, while eight (10%) were published in a general or other domain science venue. Five items (6%) were published in other venues, including magazines and newsletters. We note that our initial query did retrieve numerous computer science research papers related to database migration, but after abstract review we determined that many of these were not relevant to this study, as they focused on topics like “software analytics toolset development” (Dueñas et al., 2021), or the description of domain-specific databases without any discussion of sustainability or curation (Tabakmakher et al., 2019).

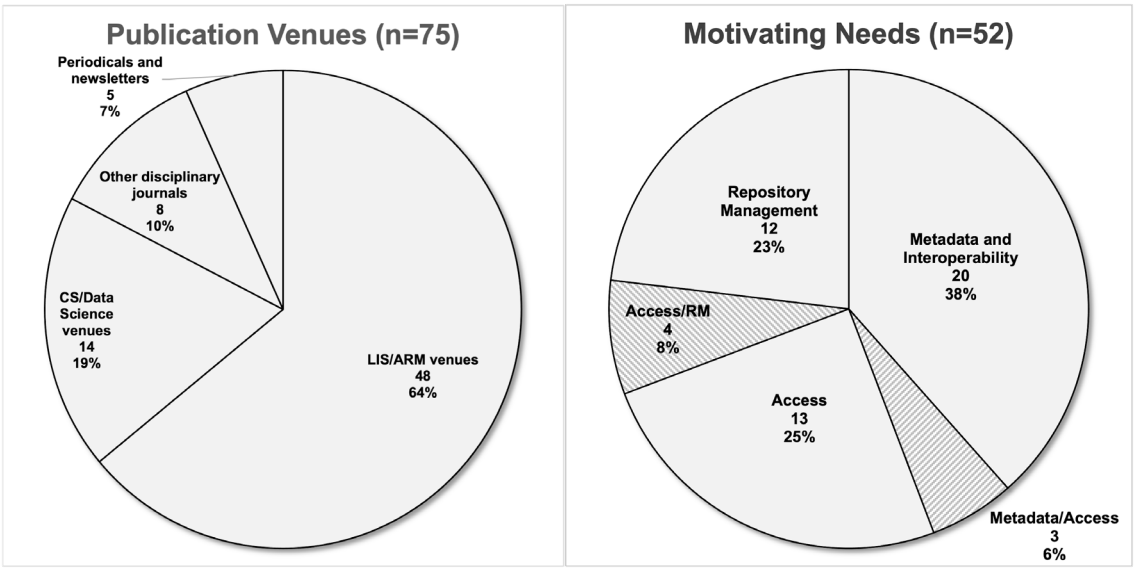


Figure 1. Left: Publication venues, showing which disciplinary venues publish work in this area. Right: Motivating needs, showing the percentages of papers discussing types of user needs as motivations for their work. We note that only 52 of the 75 papers we reviewed described motivating needs.

Similarly, most papers described database curation work taking place within an academic library (n=16) or institutional repository (n=23). An additional 17 described work in scientific data repositories. 13 papers described the curation or sustainability of individual disciplinary databases (e.g. a curated archaeology database, or a biomedical database on a specific topic, rather than a generic scientific data repository). Finally, three papers described database work in natural history museums; one paper described work in an archive; and two described database curation work in a public library.

We additionally categorized papers according to their focus, and found that while many papers discuss sustainability broadly, few specifically discuss database migration:

- 17 papers in the corpus were primarily about migration, either in the form of case studies of a particular database migration or research about migration (Allen, 2017; Bruns et al., 2014; Choi, 2020; Davis, 2015; Doty et al., 2015; Dunn et al., 2016; Fabian, 2006; Fallaw et al., 2021; Klump et al., 2015; Matusiak et al., 2017; Murphy, 2003; Pierce, 2019; Shepard, 2013; Thakar & Szalay, 2010; Thomer et al., 2017, 2018; Trippel & Zinn, 2021).
- Six focused on database maintenance in preparation for migration (Darcovich et al., 2018; Dressler, 2014; Georgieva and Flinchbaugh, 2021; Walsh, 2010; Wu et al., 2016; Wu et al., 2019).
- 39 papers were about maintenance or sustainability of a system separate from migration (Bond, 2006; Dal Pra et al., 2019; Downs & Chen, 2010; Gordon et al., 2015; Greenberg et al., 2009; Greig & Nixon, 2007; Gries et al., 2018; Han, 2011; Herrera, 2007; Imker, 2020; Jeng et al., 2017; Jiang et al., 2015; Jones et al., 2000; Kertesz et al., 2020; Lafferty-Hess et al., 2020; Le et al., 2015; Lee et al., 2018; Lutz & Meadow, 2006; Madsen & Oleen, 2013; Manninen, 2018; Marcial & Hemminger, 2010; Nagendra et al., 2001; Nie et al., 2021; Owen & Michalak, 2015; Park et al., 2018; Philipson, 2020; Post et al., 2019; Powell & Moseley, 2021; Reilly, 2009; Rossi & Ahmed, 2016; Rousidis et al., 2014; Schuler & Kesselman, 2018; Stolte et al., 2003; Trisovic et al., 2020; Tuan et al., 2011; Waldron, 2017; Wilkinson et al., 2017; Xu et al., 2016; Zohner et al., 2019).
- Two papers took a broad look at both migration and maintenance (Stigler & Steiner, 2018; Thomer & Twidale, 2014).
- And finally, 11 papers discuss migration as a background or tangential concern (Breeding, 2002; Brush & Jiras, 2019; Byrne, 2014; Gentry et al., 2021; Kansa, 2005; Kipnis MSI et al., 2019; Knight-Davis et al., 2015; Liu & Zhou, 2011; Rieh et al., 2008; Rosenthaler et al., 2015; Yin et al., 2020).

Drivers of maintenance work and migration

Most fundamentally, database maintenance and migration are motivated by a need to propagate data over time; as Breeding writes, "the long-term preservation of digital material requires a commitment to an ongoing set of processes to move digitized materials through each generation of technology" (2002). However, we found that

migration projects are additionally motivated by a desire to respond to changing user needs—whether the needs of scientific and scholarly users accessing the data, or the needs of repository and database managers working with data. These 52 instances fell into three categories: metadata and interoperability; repository management needs; and access needs more broadly (Figure 1, right).

Metadata and interoperability improvements

23 papers described efforts to improve metadata or otherwise increase the interoperability of a system to support different changing user needs. Metadata, especially legacy metadata from a previous system, was routinely described as an obstacle to the use of a digital collection when the metadata is messy and nonstandard (Darcovich et al., 2018; Georgieva & Flinchbaugh, 2021; Gries et al., 2018; Kertesz et al., 2020; Rousidis et al., 2014; Waldron, 2017), inflexible (Allen et al., 2017; Fabian, 2006; Fallaw et al., 2021; Lutz and Meadow, 2006; Mannien, 2018; Shepard, 2013), idiosyncratic (Murphy, 2003; Wilkinson et al., 2017), inappropriate for the subject (Greenberg et al., 2009; Rousidis et al., 2014), or incompatible with the hardware and software environment (Darcovich et al., 2018; Klump et al., 2015). Metadata improvements were needed to improve user search experience (Darcovich et al., 2018; Dressler, 2016; Herrera, 2007; Jiang et al., 2015; Lee et al., 2018; Murphy, 2003; Owen & Michalak, 2015; Stolte et al., 2003); support multiple search functionalities (Jiang et al., 2015; Stolte et al., 2003); incorporate metadata from legacy material (Bond, 2006; Waldron, 2017; Zohner et al., 2019); add additional information to make the metadata rich and consistent (Georgieva & Flinchbaugh, 2021; Greenberg et al., 2009; Manninen, 2018; Philipson, 2020; Rousidis et al., 2014; Waldron, 2017; Wilkinson et al., 2017); and to improve interoperability between systems (Philipson, 2020; Powell & Moseley, 2021; Wu et al., 2016, 2019). Interoperability was a particular concern for those working with domain-specific databases, where there is a need to share or aggregate data with other related systems (Gries et al., 2018; Stigler & Steiner, 2018; Stolte et al., 2003; Wilkinson et al., 2017).

Access improvements

Migrations were also motivated by a desire to centralize data to a single access point. This includes centralizing domain specific research (i.e. natural science data, or digital humanity collections) from multiple institutions and researchers (Le et al., 2015; Nagendra et al., 2001; Park et al., 2018; Rosenthaler et al., 2015; Tuan et al., 2011), adding legacy data to a new system (Dunn et al., 2016; Park et al., 2018), and combining multiple sets of digital material to a single point of access (Allen, 2017; Fabian, 2006; Lutz & Meadow, 2006; Xu et al., 2016).

Also related to access was the need of users to be able to access data remotely. Multiple authors described migration to move data to "the cloud" or through the internet (Choi, 2020; Fallaw et al., 2021; Klump et al., 2015; Shepard, 2013; Wu et al., 2016). Shepard et al. spoke of the fact that prior to a migration, users had to come into their medical archive to look at the database, and that this wasn't feasible for many researchers. Others spoke of wanting data available online so users could download and interact with data in creative ways (Rossi & Ahmed, 2016; Stolte et al., 2003).

Repository management improvements

The third category of need relates to the needs of those managing the data. These needs included a desire for more cost effective data storage (Fallaw et al., 2021; Shepard, 2013; Thakar & Szalay, 2010), efficiency in data upload and management (Allen, 2017; Walsh, 2010), and the ability to 'scale up' a current workflow (Madsen & Oleen, 2013; Wu et al., 2016, 2019). Other needs related to wanting their system to be more reliable and trustworthy to users and other archives. (Bruns et al., 2014; Gordon et al., 2015; Powell & Moseley, 2021; Rosenthaler et al., 2015). This included wanting to be Open Archives Initiative (OAI) compliant (Greig and Nixon, 2007), and following the FAIR guidelines for reproducibility (Trisovic, 2020). These guidelines most likely mean more to other repositories than users but are nevertheless notable goals to meet during a migration. This category of needs will ultimately benefit users, but align more with the operations needs of managing a repository and its respective data.

Common themes and lessons learned

Many of the papers—particularly the case studies—we reviewed describe "lessons learned" or made prescriptive recommendations for others engaging in database migration or maintenance. Below we describe some of the major themes in these, including: schema and metadata considerations; data quality and cleaning; the role of collaboration and cooperation; challengers in choosing a database system; and unexpected hurdles.

Schema and metadata considerations

A number of articles described challenges working with metadata and suggested ways of building more resilient data architectures, and recommendations for how future database curators should consider the underlying schemas in their systems. Recommendations include:

- *Consider the impact of a database's structure on end users' ability to do scholarly work with data*. Kansa in his paper on archaeological data integration notes that databases have "great theoretical significance" for the fields

that use them, and that "constraints imposed by the data schema can constrain the observational and interpretive process… database design is therefore an integral aspect of archaeological methodology, since design choices both reflect and help determine the interpretive process" (2005). Although this was only noted by one paper, we highlight this recommendation because it echoes findings from other work on infrastructure (Hilgartner, 1995; Hine, 2006).

- *Build editable systems.* Kansa also notes that "Domain experts need the ability to author, debate, and revise data mappings since some mappings will be contested" (2005). Schiller and Kesselman (2018) similarly argue that "Scientists need a semantically higher-level framework for schema evolution (i.e., transformation of schema and instance data) to evolve scientific databases more effectively and efficiently than with conventional tools or related approaches."
- *Store digital objects and metadata separately.* Klump et al. (2015) and Stolte et al. (2003) both recommend that data be stored separately from metadata so that metadata might be managed by the database, whereas files might reside in file archives. Stolte et al. note that this makes maintenance and migration of the system easier over time.
- *Use open standards.* Breeding (2002), Kansa (2005) and Shepard (2013) all recommend the use of open standards for data and metadata; Stigler and Steiner (2018) further recommend that database curators maintain a list of preferred formats for deposit.
- *Preserve legacy metadata.* Liu and Zhou (2011) recommend preserving legacy metadata as much as possible (in their case, MARC records), and note that it reduces the cost of metadata creation. Trippel and Zinn, however, warn that curators should take care when migrating data from a domain-specific to generic platform, because meaning can be lost when migrating to a more generalized schema (2021).

Data quality and cleaning

In multiple papers, authors wrote that before migration could occur, dataset- or object-level curation had to be performed, including metadata cleanup or remediation (Darcovich et al., 2018; Georgieva and Flinchbaugh, 2021; Gentry et al., 2021; Waldron, 2017; Walsh, 2010), data normalization (Georgieva and Flinchbaugh, 2021), file format migration (Allen, 2017; Dunn et al., 2016; Jeng et al., 2017), file format risk assessment (Lafferty-Hess et al., 2020), file segmentation (Matusiak et al., 2017), and checking for file corruption (Doty et al., 2015). Two papers included explicit recommendations for how to plan for this work:

- *Build quality control into multiple steps of the procedure*. This prevents mistakes from compounding and worsening over the course of a migration (Park, 2018).
- *Account for an extensive testing and transition period*. Trippel and Zinn (2021) recommend building both human-level (e.g. manual review and spot checks) and automated quality control tests (e.g. scripts that ensure that identifiers resolve as intended) after a system has been migrated.

The role of collaboration and cooperation

Many papers described the importance of working as a curatorial team, whether within one institution or as part of a broader community; database curation is not a job for a single person, department, or area of expertise.

- *Involve all departments of your institution*. Bruns (2014) states that, "In the Eastern Illinois University experience, it quickly became clear that involving the entire library was not only beneficial but a necessity for launching this important project". In their case, this meant a need for collaboration between traditional library expertise, and more technical components of a migration. Brush and Jiras (2019), Davis (2015) and Madsen and Oleen (2015) noted a similar experience of relying on multiple areas of expertise within their library or institutional repository. Some also recommended consultation with the IT department as a first step in a migration project (Byrne, 2014; Fallaw, 2021).
- *Draw on the expertise of 'disciplinary faculty'.* Knight-Davis et al. (2015) note that their successful migration would not have been successful without collaborating with scholars working with the data they migrated. They call for this especially for datasets outside of the library and information science. Fabian (2006) calls for working with those who are 'expert partners' to the data. Gordon et al. suggest that cooperation with "researchers in the target domain" engenders a more useful system (2015). Several others (Allen, 2017; Darcovich et al., 2018; Doty et al., 2015; Pierce, 2019; Post et al., 2019; Thakar and Szalay, 2010) also laude the positive impacts of collaboration between multiple stakeholders and specialists.
- *Consider multi-institutional collaboration.* This can be to support interoperability between institutions and filling knowledge gaps. Lafferty-Hess et al. describe collaboration on data curation activities between Duke University and University of North Carolina at Chapel Hill. They state, "this exercise highlighted how

cooperative models can potentially help to address knowledge gaps" (2020). This might also entail pooling financial resources. Bruns et al. emphasize that multi-institutional collaboration is desirable, especially as a "cost containment strategy" (2014). Shepard (2013) also noted the benefit of pooling financial resources between institutions.

Budgeting for database platforms

Several authors had recommendations for choosing software when migrating to a new system—and for accounting for the cost of this new system.

- *Consider the full cost of digital curation infrastructures.* Migrations and maintenance of digital systems require continued funding. Marcial and Hemminger warn us that "[Scientific Data Repositories] without substantial investment in infrastructure and support do not survive and thrive" (2010). Further, curators must understand the entire price of these systems, as there can often be extra fees associated with tasks like data cleaning, and staff training. Herrera reminds us that "It is important to determine what hidden costs are involved and if you have the resources to support all of the costs" (2007). Choi adds that the time needed to implement a new system should be considered alongside price (2020).
- *Consider whether vendored vs. open-source solutions are a fit for a specific institution*. Authors found success for both vendored (for profit) and open-source infrastructures. Brush and Jiras describe their experience with the vendored platform Digital Commons, and found that, "developing an IR using a vendor platform that has already been designed, with both user-facing and back-end functionality in place, clearly saved the authors hundreds of hours of work" (2019). However, another author describes her experience of migrating *out* of Digital Commons into the open-source platform Samvera. Excited about the decision, she states, "An open-source product can be dynamic and rewarding, but it requires ample local programming knowledge and a strong spirit of collaboration" (Pierce 2019).

Unexpected hurdles

Finally, several authors warned of unexpected obstacles to completing migrations or maintenance work.

- *Be aware that migrations may take longer or become more complex than anticipated.* Several authors noted there were certain aspects that took longer, required more planning, or were more complex than originally anticipated. This included mapping and crosswalking metadata and data schemas (Georgieva and Flinchbaugh, 2021; Gordon, 2015; Kansa, 2005; Lutz and Meadow, 2006; Matusiak, 2017; Owen and Michalak, 2015), unexpected data cleaning (Dunn, 2016; Knight-Davis et al., 2015), and integrating legacy data (Kansa, 2005; Knight-Davis et al., 2015). Several noted that there simply isn't a one size fits all solution to migrations, and that it's necessary to plan for building custom workflows (Lafferty Hess et al., 2020; Mannien, 2018; Post, 2019).
- *Be prepared for many smaller issues that may materialize well after the actual migration*. Trippel and Zinn found that despite extensive testing, many minor issues arose long after the launch of their database (2021).

DISCUSSION

The review above lays a baseline for further research on memory institution database migration and maintenance. To summarize: we found that the existing research in this area is dominated by case studies largely from academic libraries and institutional repositories; and that there are relatively few papers specifically on database and infrastructure migration, but more on maintenance and sustainability broadly. Drivers of database curation actions included technical obsolescence but also changing user needs; these were met through metadata, access, and repository management improvements. Finally, researchers had a range of "lessons learned" and suggestions for others taking on this work, notably on the topics of collaboration, data cleaning, choosing a system, and building a schema overall.

Our review reveals the following gaps that must be addressed to truly support long-term database curation going forward.

Better understanding changing user needs of infrastructure over time

Many authors described migration and maintenance actions as necessary to support different users' needs—either the scholars' who were depositing and accessing digital objects, or the database curators' themselves. In some ways, this is not surprising; existing digital curation guidelines already underscore the importance of taking a use-based approach in curating digital objects. The OAIS framework, for instance, outlines the importance of prioritizing curatorial work around the needs of a "designated community," and the DCC curation lifecycle model includes "community watch and participation" as a core curatorial activity. However, supporting user needs of an *infrastructure* is quite different from supporting user needs for a particular dataset or digital object. Research on data practices must be extended to include users' data*base* practices—and further, must consider the needs of the curators

of those databases and other information systems. The consequences of ignoring these needs include the many obstacles described in the papers we reviewed, including loss of data, interruptions in workflows, security risks, and so on.

Better financial support for, and institutional commitment to, database curation

Many papers described financial constraints to their selection of a data system, or in finding sustained staffing to oversee a migration. Additionally, they described struggles in finding sustained institutional commitment to database curation. We argue that institutions need to see database curation as a core part of providing digital curation services and infrastructure. We suspect that there are two core obstacles blocking sustained commitment and support for database curation: first, infrastructure's noted tendency to fade into the background (and therefore out of the budget) unless it is broken (Bowker & Star, 2000); and second, a persistent (though misguided) preference for funding new systems and tools rather than maintaining infrastructures that already exist (Russell & Vinsel, 2018).

One way of forcing the important work of database curation into more persistent view might be to better represent it in high-level models of digital curation like the IDCC lifecycle model, or OAIS. So many of our digital curation models focus on the maintenance of individual digital objects but assume a stable infrastructure in which to store them (or their metadata). In our future work, we hope to model the rhythms of database maintenance and migration in LAMs. Though we suspect there is not a strict seasonality to migrating databases, there are common motivators and pressures that require migrations, such as changing user needs or software obsolescence. Modeling these pressures may make it easier to plan and fund database curation in the long-term.

The impact of the rigidity of data architecture on sustainability

Many authors described obstacles to database curation that essentially come down to issues with data architecture and structure: schemas that are hard to change; legacy metadata that resists incorporation into new systems; and data that requires extensive cleaning before aggregation. We consequently ask: might be done to make databases more usable to their administrators and curators, and more manageable over time? In our prior work (Thomer et al., 2018; 2020) we argued that better tools and interfaces are needed to support database curators in managing systems over time; here we specifically see a need for tools that better support editing or evolving the underlying schemas of data systems, and that facilitate bulk transformations of metadata records as they're moved and aggregated from one system to another. Some of the research we reviewed out of the computer science literature hints at possibilities for these tools and interfaces. However, significant collaboration between database curators and computer scientists would be needed to make such tools truly feasible.

Building more intentional communities of (database) practice

Finally, we note that one of the most common "lessons learned" in the papers reviewed was the role that community and cooperation plays in facilitating database curation. This includes collaboration across departments, institutions, and in some cases fields. In our prior work, we have similarly found the importance that community plays in supporting practitioners as they work through the complex issues of a database migration (Thomer et al., 2018; Rayburn & Thomer 2022). We believe that there is a need to build a more intentional community of practice around database curation—including database curators engaged in maintenance work, computer scientists exploring modes of facilitating schema evolution and tooling, and the scholars that rely on these infrastructures as archives and access points.

CONCLUSION

In this paper, we have conducted an integrative review of computer, library and information sciences research on the topic of database migration, maintenance and sustainability. The work of database curation is crucial for the longevity of our digital collections and infrastructures yet is not well supported by existing best practices or frameworks. Existing research in this area is dominated by individual case studies, indicating a need for more synthesis and generalizable research. That said, by synthesizing case study findings, we were able to identify recurring challenges around schema crosswalking, data cleaning, and budgeting, and a common reliance on community support to surmount challenges. We argue that there is a need for better methods of understanding changing infrastructure practices over time (rather than just data practices). We also need to account for the work of infrastructure maintenance in digital curation frameworks, and thereby bring more visibility (and possibly, more funding and institutional support) to this critical work. We additionally call for a stronger community of practice around database curation, bringing together curators, scholars, and computer science researchers interested in building more changeable and user-friendly data systems. Going forward, we intend to continue our work modeling the rhythms of database migration in memory institutions and bridging those with existing digital curation lifecycle models. In doing so, we hope to provide practitioners with greater guidance and support for this work, and to contribute to theory on infrastructural maintenance and change.

ACKNOWLEDGMENTS

This research was funded by IMLS grant RE-07-18-0118-18. The authors thank Rebecca A. Welzenbach (Research Impact and Information Science Librarian at the University of Michigan) for her assistance in selecting databases, creating search queries, and framing the literature search. The bibliography and codes underlying this literature review can be found at https://doi.org/10.7302/5hr4-j779.

REFERENCES

Allen, A. L. (2017). Lessons Learned in Partnerships and Practice: Adopting Open Source Institutional Repository Software. *Journal of Librarianship and Scholarly Communication*, *5*. http://dx.doi.org.proxy.lib.umich.edu/10.7710/2162-3309.2170

Bond, T. J. (2006). Sustaining a digital collection after the grants: The early Washington maps project. *OCLC Systems and Services*, *22*(1), 56–66. Scopus. https://doi.org/10.1108/10650750610640810

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences* (First paperback edition). The MIT Press.

Breeding, M. (2002, May). Preserving digital information: Building collections that will outlast current technologies is a challenge. (The Systems Librarian). *Information Today*, *19*(5), 48+. Gale OneFile: Computer Science.

Brodie, M. L., & Stonebraker, M. (1995). *Migrating legacy systems: Gateways, interfaces & the incremental approach*. Morgan Kaufmann Publishers ; IT/Information Technology.

Bruns, T. A., Knight-Davis, S., Corrigan, E. K., & Brantley, S. (2014). It Takes a Library: Growing a Robust Institutional Repository in Two Years. *College & Undergraduate Libraries*, *21*(3–4), 244–262. http://dx.doi.org.proxy.lib.umich.edu/10.1080/10691316.2014.904207

Brush, D. A., & Jiras, J. (2019). Developing an institutional repository using Digital Commons. *Digital Library Perspectives*, *35*(1), 31–40. Library & Information Science Abstracts (LISA). https://doi.org/10.1108/DLP-08-2017-0028

Buneman, P., Cheney, J., Tan, W.-C., & Vansummeren, S. (2008). Curated Databases. *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 1–12. https://doi.org/10.1145/1376916.1376918

Byrne, E. J. (2014). CONTENTdm and Content Pro: A Comparison and Evaluation. *Library Philosophy and Practice*, 1–19.

Choi, C. Q. (2020). Migrating big astronomy data to the cloud. *Nature*, *584*(7819), 159–160. Scopus. https://doi.org/10.1038/d41586-020-02284-7

Consultative Committee for Space Data Systems. (2012). *Reference Model for an Open Archival Information System (OAIS)* (650.0-M–2). https://public.ccsds.org/pubs/650x0m2.pdf

Cook, R., Michener, W. K., Vieglais, D. A., Budden, A. E., & Koskela, R. J. (2012). DataONE: A Distributed Environmental and Earth Science Data Network Supporting the Full Data Life Cycle. In A. Abbasi & N. Giesen (Eds.), *EGU General Assembly Conference Abstracts*.

Dal Pra, S., Falabella, A., Fattibene, E., Cincinelli, G., Magnani, M., De Cristofaro, T., & Ruini, M. (2019). Evolution of monitoring, accounting and alerting services at INFN-CNAF Tier-1. In A. Forti, L. Betev, M. Litmaath, O. Smirnova, & P. Hristov (Eds.), *23rd International Conference on Computing in High Energy and Nuclear Physics (CHEP 2018), 9-13 July 2018* (Vol. 214, p. 08033). EDP Sciences. https://doi.org/10.1051/epjconf/201921408033

Darcovich, J., Flynn, K., & Mingyan Li. (2018). Born of Collaboration: The Evolution of Metadata Standards in an Aggregated Environment. *Visual Resources Association Bulletin*, *45*(2), 1–12. Library & Information Science Source.

Davis, L. D. (2015). Migrating The Breeze: A Case Study In Cooperative Staffing. *Journal for the Society of North Carolina Archivists*, *12*(1 & 2), 41–52. Library & Information Science Source.

Doty, J., Kowalski, M. T., Nash, B. C., & O'Riordan, S. F. (2015). Making Student Research Data Discoverable: A Pilot Program Using Dataverse. *Journal of Librarianship and Scholarly Communication*, *3*(2). http://dx.doi.org.proxy.lib.umich.edu/10.7710/2162-3309.1234

Downs, R. R., & Chen, R. S. (2010). Self-Assessment of a Long-Term Archive for Interdisciplinary Scientific Data as a Trustworthy Digital Repository. *Journal of Digital Information*, *11*(1). Library & Information Science Abstracts (LISA). https://proxy.lib.umich.edu/login?url=https://www.proquest.com/scholarly-journals/self-assessment-long-term-archive/docview/742899171/se-2?accountid=14667

Dressler, V. A. (2014). Re-platforming digital collections for enhanced access & search functionality. *Proceedings of the Association for Information Science & Technology*, *51*(1), 213–215. Library & Information Science Source.

Dressler, V. A. (2016). Investigating and implementing an extensible, adaptable game plan for digital initiatives at a large state university. *The Electronic Library*, *34*(4), 588–596. Library & Information Science Abstracts (LISA). https://doi.org/10.1108/EL-02-2015-0034

Dueñas, S., Cosentino, V., Gonzalez-Barahona, J. M., Felix, A. del C. S., Izquierdo-Cortazar, D., Cañas-Díaz, L., & García-Plaza, A. P. (2021). GrimoireLab: A toolset for software development analytics. *PeerJ Computer Science*, *7*, e601. https://doi.org/10.7717/peerj-cs.601

Dunn, W. D., Cobb, J., Levey, A. I., & Gutman, D. A. (2016). REDLetr: Workflow and tools to support the migration of legacy clinical data capture systems to REDCap. *International Journal of Medical Informatics*, *93*, 103–110. https://doi.org/10.1016/j.ijmedinf.2016.06.015

Fabian, C. A. (2006). UBdigit: A repository infrastructure for digital collections at the University at Buffalo. *RLG DigiNews*, *10*(3), np.

Fallaw, C., Schmitt, G., Luong, H., Colwell, J., & Strutz, J. (2021). Institutional Data Repository Development, a Moving Target. *Code4Lib Journal*, *51*, N.PAG-N.PAG. Library & Information Science Source.

Gentry, L. M., Ryan, E., Rayman, J., & Bace, M. (2021). How to Wrangle Multiple Discrete Collections from One Donor: A Case Study of the Subject-based Physical and Digital Consolidation of the Wade Hall Collections. *American Archivist*, *84*(1), 62–90. Library Literature & Information Science Full Text (H.W. Wilson). https://doi.org/10.17723/0360-9081-84.1.62

Georgieva, M., & Flinchbaugh, M. (2021). Biz of Digital - Metadata Remediation of Legacy Digital Collections: Efficient Large-Scale Metadata Clean-Up with a Sleek Workflow and a Handy Tool. *Against the Grain*, *33*(1), 47.

Gordon, A. S., Millman, D. S., Steiger, L., Adolph, K. E., & Gilmore, R. O. (2015). Researcher-Library Collaborations: Data Repositories as a Service for Researchers. *Journal of Librarianship and Scholarly Communication*, *3*(2). https://doi.org/10.7710/2162-3309.1238

Greenberg, J., White, H. C., Carrier, S., & Scherle, R. (2009). A Metadata Best Practice for a Scientific Data Repository. *Journal of Library Metadata*, *9*(3–4), 194–212. Library & Information Science Abstracts (LISA). https://doi.org/10.1080/19386380903405090

Greig, M., & Nixon, W. J. (2007). On the road to Enlighten-ment: Establishing an institutional repository service for the University of Glasgow. *International Digital Library Perspectives*, *23*(3), 297–309. Library & Information Science Abstracts (LISA). https://doi.org/10.1108/10650750710776431

Gries, C., Budden, A., Laney, C., O'brien, M., Servilla, M., Sheldon, W., Vanderbilt, K., & Vieglais, D. (2018). Facilitating and improving environmental research data repository interoperability. *Data Science Journal*, *17*. Scopus. https://doi.org/10.5334/dsj-2018-022

Han, Y. (2011). Cloud Computing: Case Studies and Total Cost of Ownership. *Information Technology and Libraries*, *30*(4), 198–206. https://doi.org/10.6017/ital.v30i4.1871

Herrera, G. (2007). MetaSearching and Beyond: Implementation Experiences and Advice from an Academic Library. *Information Technology and Libraries*. https://egrove.olemiss.edu/libpubs/14

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, *3*(1), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

Hilgartner, S. (1995). Biomolecular Databases: New Communication Regimes for Biology? *Science Communication*, *17*(2), 240–263. https://doi.org/10.1177/1075547095017002009

Hine, C. (2006). Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science*, *36*(2), 269–298. https://doi.org/10.1177/0306312706054047

Imker, H. J. (2020). Who Bears the Burden of Long-Lived Molecular Biology Databases? *Data Science Journal*, *19*(1), 8. https://doi.org/10.5334/dsj-2020-008

Jackson, S. J. (2014). Rethinking Repair. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies* (pp. 221–240). The MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0011

Jagadish, H. V., Chapman, A., Elkiss, A., Jayapandian, M., Li, Y., Nandi, A., & Yu, C. (2007). *Making database systems usable*. 13. https://doi.org/10.1145/1247480.1247483

Jeng, W., He, D., & Chi, Y. (2017). Social science data repositories in data deluge. *Electronic Library*, *35*(4), 626–649. Library, Information Science & Technology Abstracts.

Jiang, L., Zhao, Y., Xu, B., & Wen, H. (2015). The Development of a Combined Search for a Heterogeneous Chemistry Database. *Data Science Journal*, *14*(0), 3. https://doi.org/10.5334/dsj-2015-003

Jones, A. C., Sutherland, I., Embury, S. M., Gray, W. A., White, R. J., Robinson, J. S., Bisby, F. A., & Brandt, S. M. (2000). Techniques for effective integration, maintenance and evolution of species databases. In O. Gunther & H. J. Lenz (Eds.), *12th International Conference on Scientific and Statistical Database Management, Proceedings* (pp. 3–13). Ieee Computer Soc. https://doi.org/10.1109/SSDM.2000.869774

Kansa, E. (2005). A community approach to data integration: Authorship and building meaningful links across diverse archaeological data sets. *Geosphere*, *1*(2), 97–109. https://doi.org/10.1130/GES00013.1

Kertesz, V., Monguidi, M., & Pasinato, L. (2020). Database on apple fruit pests of the EU to support pest risk assessments. *Efsa Journal*, *18*(5), 6149. https://doi.org/10.2903/j.efsa.2020.6149

Kipnis MSI, D. G., Palmer, L. A., & Kubilius, R. K. (2019). The institutional repository landscape in medical schools and academic health centers: A 2018 snapshot view and analysis. *Journal of the Medical Library Association*, *107*(4), 488–498. http://dx.doi.org.proxy.lib.umich.edu/10.5195/jmla.2019.653

Klump, J., Ulbricht, D., & Conze, R. (2015). Curating the web's deep past—Migration strategies for the German Continental Deep Drilling Program web content. *GeoResJ*, *6*, 98–105. https://doi.org/10.1016/j.grj.2015.02.011

Knight-Davis, S., Bruns, T., & Tucker, G. C. (2015). Big Things Have Small Beginnings: Curating a Large Natural History Collection-Processes and Lessons Learned. *Journal of Librarianship & Scholarly Communication*, *3*(2), 1–21. Library & Information Science Source.

Lafferty-Hess, S., Rudder, J., Downey, M., Ivey, S., Darragh, J., & Kati, R. (2020). Conceptualizing Data Curation Activities Within Two Academic Libraries. *Journal of Librarianship & Scholarly Communication*, *8*, 1–16. Library & Information Science Source.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Le, V. D., Neff, M. M., Stewart, R. V., Kelley, R., Fritzinger, E., Dascalu, S. M., & Harris, F. C. (2015). Microservice-based architecture for the NRDC. *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*, 1659–1664. https://doi.org/10.1109/INDIN.2015.7281983

Lee, R. Y. N., Howe, K. L., Harris, T. W., Arnaboldi, V., Cain, S., Chan, J., Chen, W. J., Davis, P., Gao, S., Grove, C., Kishore, R., Muller, H.-M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., Rodgers, F., Russell, M., Schindelman, G., … Sternberg, P. W. (2018). WormBase 2017: Molting into a new stage. *Nucleic Acids Research*, *46*(D1), D869–D874. https://doi.org/10.1093/nar/gkx998

Liu, S., & Zhou, Y. (2011). Developing an institutional repository using DigiTool. *The Electronic Library*, *29*(5), 589–608. https://doi.org/10.1108/02640471111177044

Lutz, M., & Meadow, C. (2006). Evolving an in-house system to integrate the management of digital collections. *Library Hi Tech*, *24*(2), 241. Library & Information Science Abstracts (LISA). https://doi.org/10.1108/07378830610669619

Madsen, D. L., & Oleen, J. K. (2013). Staffing and Workflow of a Maturing Institutional Repository. *Journal of Librarianship and Scholarly Communication*, *1*(3). Library & Information Science Abstracts (LISA). https://doi.org/10.7710/2162-3309.1063

Manninen, L. (2018). Describing Data: A Review of Metadata for Datasets in the Digital Commons Institutional Repository Platform: Problems and Recommendations. *Journal of Library Metadata*, *18*(1), 1–11. Library & Information Science Abstracts (LISA). https://doi.org/10.1080/19386389.2018.1454379

Marcial, L. H., & Hemminger, B. M. (2010). Scientific data repositories on the Web: An initial survey. *Journal of the American Society for Information Science and Technology*, *61*(10), 2029–2048. Scopus. https://doi.org/10.1002/asi.21339

Matusiak, K. K., Tyler, A., Newton, C., & Polepeddi, P. (2017). Finding access and digital preservation solutions for a digitized oral history project: A case study. *Digital Library Perspectives*, *33*(2), 88–99. Scopus. https://doi.org/10.1108/DLP-07-2016-0025

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–23. https://doi.org/10.1145/3359174

Murphy, J. L. (2003). Link It or Lump It: Basic Access Strategies for Digital Art Representation. *Journal of Library Administration*, *39*(2/3), 139–160. Library & Information Science Source.

Nagendra, K. S., Bukhres, O., Sikkupparbathyam, S., Areal, M., Miled, Z. B., Olsen, L., Gokey, C., Kendig, D., Northcutt, T., Kordova, R., & Major, G. (2001). NASA Global Change Master Directory: An implementation of asynchronous management protocol in a heterogeneous distributed environment. *Proceedings 3rd International Symposium on Distributed Objects and Applications*, 136–145. https://doi.org/10.1109/DOA.2001.954079

Nie, H., Pengcheng Luo, & Ping Fu. (2021). Research Data Management Implementation at Peking University Library: Foster and Promote Open Science and Open Data. *Data Intelligence*, *3*(1), 189–204. https://doi.org/10.1162/dint_a_00088

Okoli, C., & Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1954824

Owen, W., & Michalak, S. C. (2015). Engine Of Innovation: Building the High Performance Catalog. *Information Technology and Libraries*, *34*(2), 5–18. https://doi.org/10.6017/ital.v34i2.5702

Park, E. G., Burr, G., Slonosky, V., Sieber, R., & Podolsky, L. (2018). Data rescue archive weather (DRAW): Preserving the complexity of historical climate data. *Journal of Documentation*, *74*(4), 763–780. Library & Information Science Abstracts (LISA). https://doi.org/10.1108/JD-10-2017-0150

Pati, D., & Lorusso, L. N. (2018). How to Write a Systematic Review of the Literature. *HERD: Health Environments Research & Design Journal*, *11*(1), 15–30. https://doi.org/10.1177/1937586717747384

Philipson, J. (2020). The Red Queen in the Repository. *International Journal of Digital Curation*, *15*(1). https://doi.org/10.2218/ijdc.v15i1.646

Pierce, P. (2019). Biz of Digital—Transitioning to a New IR Platform. *Against the Grain*, *31*(6), 58–59. Library, Information Science & Technology Abstracts.

Post, C., Chassanoff, A., Lee, C. A., Rabkin, A., Zhang, Y., Skinner, K., & Meister, S. (2019). Digital curation at work: Modeling workflows for digital archival materials. In *Proceedings of the 18th Joint Conference on Digital Libraries* (pp. 39–48). IEEE Press. https://doi.org/10.1109/JCDL.2019.00016

Powell, C. D., & Moseley, H. N. B. (2021). The mwtab Python Library for RESTful Access and Enhanced Quality Control, Deposition, and Curation of the Metabolomics Workbench Data Repository. *Metabolites*, *11*(3), 163. https://doi.org/10.3390/metabo11030163

Rayburn, A., & Thomer, A. K. (2022, February). The Craft of Database Curation: Taking Cues from Quiltmaking. *iConference*, virtual.

Reilly, B. (2009). CRL's long-lived digital collections project: Working to provide member libraries peace-of-mind. *Against the Grain*, *21*(2), 34–36.

Ribes, D., & Finholt, T. A. (2009). "The Long Now of Technology Infrastructure: Articulating Tensions in Development. *Journal of the Association for Information Systems*, *10*(5). https://doi.org/10.17705/1jais.00199

Rieh, S. Y., St Jean, B., Yakel, E., Markey, K., & Kim, J. (2008). Perceptions and Experiences of Staff in the Planning and Implementation of Institutional Repositories. *Library Trends*, *57*(2), 168–190. Library & Information Science Abstracts (LISA). https://doi.org/10.1353/lib.0.0027

Rosenthaler, L., Fornaro, P., & Clivaz, C. (2015). DASCH: Data and Service Center for the Humanities. *Digital Scholarship in the Humanities*, *30*(suppl1), 43–49. https://doi.org/10.1093/llc/fqv051

Rossi, R. A., & Ahmed, N. K. (2016). An Interactive Data Repository with Visual Analytics. *ACM SIGKDD Explorations Newsletter*, *17*(2), 37–41. https://doi.org/10.1145/2897350.2897355

Rousidis, D., Garoufallou, E., Balatsoukas, P., & Sicilia, M.-A. (2014). Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories. *Information Services & Use*, *34*(3–4), 279–286. https://doi.org/10.3233/ISU-140746

Russell, A. L., & Vinsel, L. (2018). After Innovation, Turn to Maintenance. *Technology and Culture*, *59*(1), 1–25. https://doi.org/10.1353/tech.2018.0004

Schuler, R. E., & Kesselman, C. (2018). Towards an efficient and effective framework for the evolution of scientific databases. *Proceedings of the 30th International Conference on Scientific and Statistical Database Management*, 1–4. https://doi.org/10.1145/3221269.3221300

Shepard, E. (2013). A Digital Collection Collaboration in an Academic Repository: A Case Study. *Journal of Archival Organization*, *11*(3/4), 205–220. Library & Information Science Source. https://doi.org/10.1080/15332748.2013.948739

Stigler, J., & Steiner, E. (2018). GAMS – An infrastructure for the long-term preservation and publication of research data from the humanities. *VOEB-Mitteilungen*, *71*(1), 207–216. Scopus. https://doi.org/10.31263/voebm.v71i1.1992

Stolte, E., von Praun, C., Alonso, G., & Gross, T. (2003). Scientific data repositories: Designing for a moving target. *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 349–360. https://doi.org/10.1145/872757.872800

Tabakmakher, V. M., Krylov, N. A., Kuzmenkov, A. I., Efremov, R. G., & Vassilevski, A. A. (2019). Kalium 2.0, a comprehensive database of polypeptide ligands of potassium channels. *Scientific Data*, *6*(1), 73 (8 pp.). https://doi.org/10.1038/s41597-019-0074-x

Thakar, A., & Szalay, A. (2010). Migrating a (large) science database to the cloud. *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, 430–434. https://doi.org/10.1145/1851476.1851539

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*(7), 545–558. https://doi.org/10.5771/0943-7444-2017-7-545

Thomer, A. K., Rayburn, A. J., & Tyler, A. R. B. (2020). Three approaches to documenting database migrations. *International Journal of Digital Curation*, *15*(1), 5. https://doi.org/10.2218/ijdc.v15i1.726

Thomer, A. K., & Twidale, M. B. (2014, March 1). *How databases learn*. iConference 2014. https://doi.org/10.9776/14409

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 Association for Information Science and Technology*, *55*(1), 504–513. https://doi.org/10.1002/pra2.2018.14505501055

Thomer, A. K., & Wickett, K. M. (2020). Relational data paradigms: What do we learn by taking the materiality of databases seriously? *Big Data & Society*, *7*(1), 205395172093483. https://doi.org/10.1177/2053951720934838

Trippel, T., & Zinn, C. (2021). Lessons learned: On the challenges of migrating a research data repository from a research institution to a university library. *Language Resources and Evaluation*, *55*(1), 191–207. https://doi.org/10.1007/s10579-019-09474-4

Trisovic, A., Durbin, P., Schlatter, T., Durand, G., Barbosa, S., Brooke, D., & Crosas, M. (2020). Advancing Computational Reproducibility in the Dataverse Data Repository Platform. *Proceedings of the 3rd International Workshop on Practical Reproducible Evaluation of Computer Systems*, 15–20. https://doi.org/10.1145/3391800.3398173

Tuan, W.-J., Sheehy, A. M., & Smith, M. A. (2011). Building a diabetes screening population data repository using electronic medical records. *Journal of Diabetes Science and Technology*, *5*(3), 514–522. Scopus. https://doi.org/10.1177/193229681100500306

United States Geological Survey. (2014). *The United States Geological Survey Science Data Lifecycle Model* (USGS Numbered Series No. 2013–1265). U.S. Geological Survey. 10.3133/ofr20131265

University of Illinois Urbana-Champaign School of Information Science. (2006). *Data Curation*. University of Illinois Urbana-Champaign School of Information Science. https://ischool.illinois.edu/research/areas/data-curation

Waldron, Z. (2017). Life Story: Creating Metadata for the Portrait File. *Visual Resources Association Bulletin*, *44*(1), 1–9. Library & Information Science Source.

Walsh, M. P. (2010, September). Batch loading collections into DSpace: Using Perl scripts for automation and quality control. *Information Technology and Libraries*, *29*(3), 117+. Gale OneFile: Computer Science.

Wilkinson, M. D., Verborgh, R., Santos, L. O. B. da S., Clark, T., Swertz, M. A., Kelpin, F. D. L., Gray, A. J. G., Schultes, E. A., Mulligen, E. M. van, Ciccarese, P., Kuzniar, A., Gavai, A., Thompson, M., Kaliyaperumal, R., Bolleman, J. T., & Dumontier, M. (2017). Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Computer Science*, *3*, e110. https://doi.org/10.7717/peerj-cs.110

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, *22*(1), 45–55. https://doi.org/10.1057/ejis.2011.51

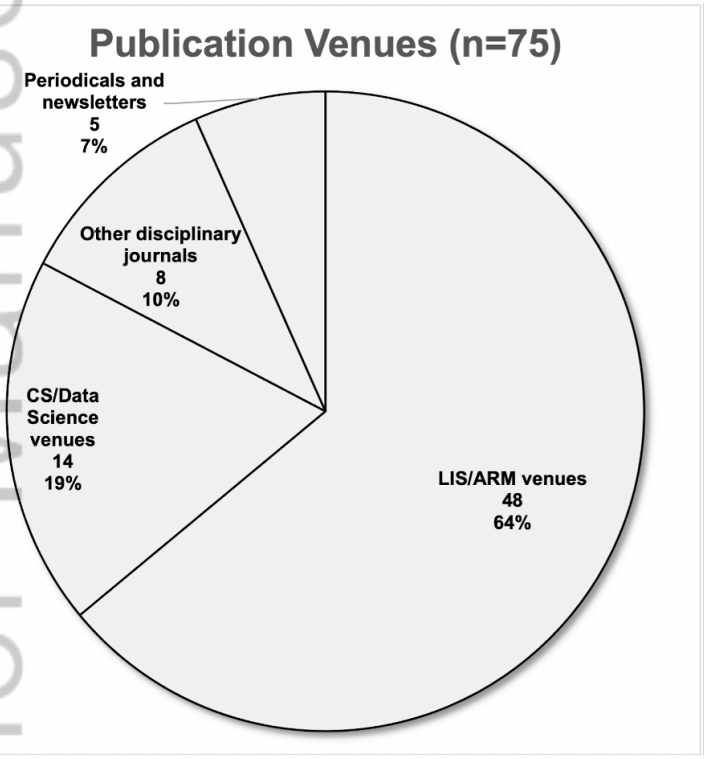
Wu, A., Davis-Van Atta, T., Thompson, S., Scott, B., Washington, A., & Xiping Liu. (2019). From Meow to ROAR: Expanding Open Access Repository Services at the University of Houston Libraries. *Journal of Librarianship and Scholarly Communication*, *7*(1), 1–20. Library & Information Science Source. https://doi.org/10.7710/2162-3309.2309

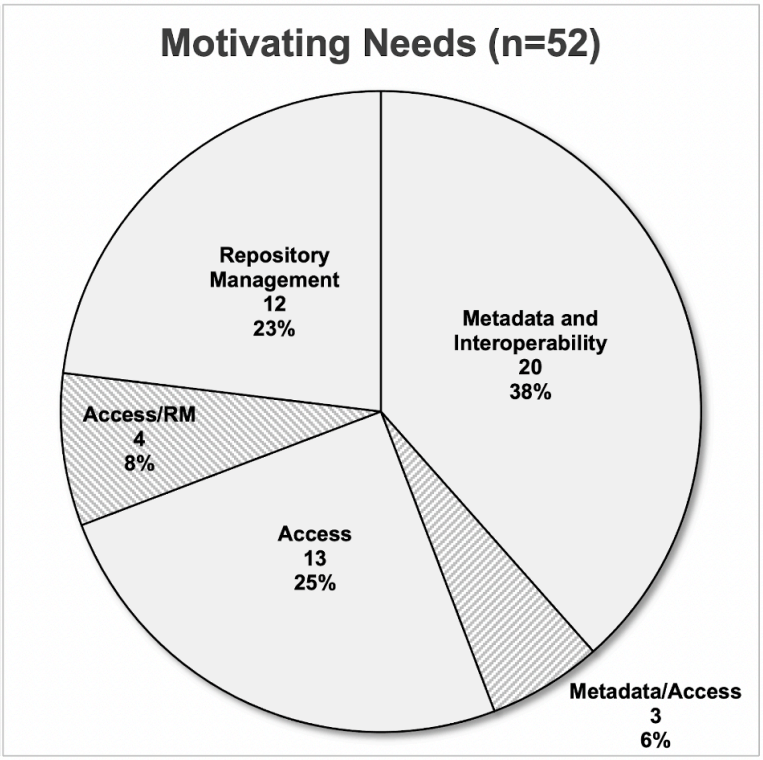
Wu, A., Thompson, S., Vacek, R., Watkins, S., & Weidner, A. (2016). Hitting the Road Towards a Greater Digital Destination: Evaluating and Testing DAMS at University of Houston Libraries. *Information Technology and Libraries*, *35*(2), 5–18. https://doi.org/10.6017/ital.v35i2.9152

Xu, W., Huang, R., Esteva, M., Song, J., & Walls, R. (2016). Content-based comparison for collections identification. *2016 IEEE International Conference on Big Data (Big Data)*, 3283–3289. https://doi.org/10.1109/BigData.2016.7840987

Yin, S., Zhang, J., Jia, M., & Hu, J. (2020). How to Evaluate and Select a Data Repository for Humanities and Social Science: A Case Study of Fudan University Data Repository for Humanities and Social Science. *Library Trends*, *69*(1), 125–137. Library & Information Science Abstracts (LISA). http://dx.doi.org.proxy.lib.umich.edu/10.1353/lib.2020.0024

Zohner, J., Marquardt, K., Schneider, H., & Backofen, A. M. (2019). Challenges and Opportunities in Changing Data Structures of Clinical Document Archives from HL7-V2 to FHIR-Based Archive Solutions. *Medinfo 2019: Health and Wellbeing E-Networks for All*, *264*





Thomer A LP22 1.png