Maintaining repositories, databases, and digital collections in memory institutions: an integrative review supplement data

**Andrea K. Thomer, Joseph R. Starks, Alexandria Rayburn, Michael Lenard**

**Data Context**

This table is meant to provide additional context for data presented in a paper published at the 85th Annual Meeting of the Association for Information Science and Technology (ASIS&T) 2022. The paper is titled, *Maintaining repositories, databases, and digital collections in memory institutions: an integrative review.*

**Data Description**

This dataset is the results from qualitatively coding the 76 articles that represent the dataset from our literature review. We 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). We also coded for the focus of the paper, and whether motivating needs were listed as a reason for migration.

**Methods**

This text is taken directly from the paper, where we describe our methods in depth.

“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 on the subject of 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” [(](https://www.zotero.org/google-docs/?GmFym3)McDonald [et al., 2019)](https://www.zotero.org/google-docs/?GmFym3), 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)](https://www.zotero.org/google-docs/?auKob0). 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 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.”

**Column heading meanings and field explanations**

Each row in the table is one of the 75 papers included in the final literature review. Below is the explanation of each column

1. Citation: A stable way to reference the literature. This is unique for each article.
2. Author's role: This is to identify the context from which the author is looking at migrations. The potential values for this field are practitioner, researcher, neither researcher nor practitioner, researcher and practitioner.
3. Publication type: This is to identify the research domain where this article was published. The potential values are CS/Data Science [ Computer Science/Data Science], LIS/ARM journal [Library Information Science, Archives Record Management journal], Domain journal, other.
4. Paper type: This was to identify the methodological approach used within the paper. The potential values are case study, research project, position paper, other.
5. Setting/domain: This was to identify the type of institution this research took place in. The potential values are Library (IR), Library (collections), Domain/scientific database, Various
6. Paper focus: This was to identify which aspect of migration the authors are focusing on within the paper. The potential values for this field are maintenance in prep for migration, maintenance or sustainability of system separate from migration, migration as background or tangential concern, primarily about migration.
7. Motivating needs: This column is for papers that expressed the motivation for migration was rooted in some unmet user needs. These could either be front end users such as researchers, or user needs for those who maintain the system. The potential values for this field are access, Access/ Repository Management, Metadata and Interoperability, Metadata and Interoperability / Access, or No user needs discussed.

**References**

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

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://www.zotero.org/google-docs/?k0Vysw) https://doi.org/10.1145/1376916.1376918

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

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

[Okoli, C., & Schabram, K](https://www.zotero.org/google-docs/?k0Vysw). (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.1954824>

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

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