Supporting the Long-term Curation and Migration of Natural History Museum Collections Databases

Andrea K. Thomer
School of Information, University of Michigan, USA.
athomer@umich.edu

Michael B. Twidale
School of Information Sciences, University of Illinois at Urbana-Champaign, USA.
twidale@illinois.edu

Nicholas M. Weber
Information School, University of Washington, USA.
nmweber@uw.edu

ABSTRACT

Migration of data collections from one platform to another is an important component of data curation – yet, there is surprisingly little guidance for information professionals faced with this task. Data migration may be particularly challenging when these data collections are housed in relational databases, due to the complex ways that data, data schemas, and relational database management software become intertwined over time. Here we present results from a study of the maintenance, evolution and migration of research databases housed in Natural History Museums. We find that database migration is an on-going – rather than occasional – process for many Collection managers, and that they creatively appropriate and innovate on many existing technologies in their migration work. This paper contributes descriptions of a preliminary set of common adaptations and "migration patterns" in the practices of database curators. It also outlines the strategies they use when facing collection-level data migration and describes the limitations of existing tools in supporting LAM and "small science" research database migration. We conclude by outlining future research directions for the maintenance and migration of collections and complex digital objects.

KEYWORDS

Data curation, database migration, natural history museums, museum informatics

INTRODUCTION

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.)) is a fundamental task in data curation and long-term digital preservation – yet, there is surprisingly little guidance for information professionals faced with this task. While metadata standards, interoperability guidelines, and careful selection of preservation-ready file formats certainly render individual digital objects more portable from one system to another, these approaches don't necessarily support the migration of databases or *collections* of data, in which the complex relationships *between* data points must be maintained. Similarly, best practices in digital preservation do not necessarily support curators in the complex tasks (such as data modeling, schema matching, data transformation, provenance capture, etc.) entailed in migrating an entire data collection or database. And while there are certainly best practices aimed at supporting enterprise-level database migrations, these approaches do not easily scale down to library, archive, museum, or "small science" scholarly contexts (Cragin, Palmer, Carlson, & Witt, 2010), in which data are more likely to be stored in idiosyncratically-structured databases, distributed over a number of legacy systems, and managed on an *ad hoc* basis by information professionals with a number of other work responsibilities. Now that libraries, archives and museums (LAMs) primarily manage their collections catalogs in digital databases, the need to develop strategies to migrate and manage databases over time has become urgent. Research is needed to understand the unique database migration needs and workflows of these communities, and to develop appropriately scaled best practices.

In the following paper, we take Natural History Museums (NHMs) as a salient use case for exploring the maintenance and migration of data collections. NHM specimen collections are a vitally important source of scientific and cultural data (NTSC, 2009). Numerous NHMs publish their specimen records through their own web-facing databases and/or through platforms such as the Global Biodiversity Information Facility (GBIF) and the System for Earth Sample Registry; and many of the databases central to modern of bio-, biodiversity and earth science informatics have their roots in NHM collections. Furthermore, as relatively early adopters of database technologies, NHMs offer a unique longitudinal view on data curation and database management, which may be informative to many other research and LAM contexts.

Our work here is part of a larger project studying the migration of research data collections and databases more generally. Our goal is to understand how collections data and databases are changed over time, and how their managers cope with, mitigate, and facilitate that change. In this broader project we ask,

1. What patterns of use, access, and obsolescence drive and support the migration of research data collections, and how?
2. What are the curatorial processes that support the on-going use of research data collections between migrations?
3. How does the use of customizable vs. standardized “off-the-shelf” database platforms impact the accessibility and maintainability of research data collections? Open access vs. closed?

By answering these questions, we can better understand on aspect of the rote but critical work of infrastructure maintenance, and thereby contribute to library and information science theory as well as the construction of more sustainable data systems.

MOTIVATION: COLLECTIONS DATABASES, DATA COLLECTIONS, AND DATA MIGRATION

This project is motivated not just by the longstanding use of relational databases in scholarly research, but also the growing use of LAM collections catalogs as datasets in and of themselves (Padilla, 2018). Card catalogs, collection ledgers, and collection databases have long been the central access point and organizational system for LAMs. However, a growing number of institutions now publish their collection data for research use beyond the physical collection (see for example Chan, 2014; “DPLA API Codex,” n.d.; Tate, 2013/2018). The use of collections catalogs as data is driving an urgent need for digital *collections* curation, in addition to the curation of individual digital objects. Not only must collections databases be maintained to ensure access to the physical materials they represent, but also to ensure the continued reliability of scholarship produced with the collection-as-dataset.

The need to curate digital collections as an entity unto themselves poses several novel challenges for the work of data curation. First, the management of aggregates, groupings, sets, and collections of digital objects requires understanding the relationship between collection- and item-level representations and metadata (Wickett, in press; Wickett, Renear, & Urban, 2010; Zavalina, Palmer, Jackson, & Han, 2009). Further, management of the infrastructure used to store collections data – often, relational databases – requires an understanding of the complex interplay between numerous entities and representation layers: the individual records in the collection; the data schemas that pull them together; the physical hard drives they are stored on; the physical objects they represent; and the uses and users of the collection; and the ways in which digital aggregates change and grow over time (Buneman, Chapman, & Cheney, 2006; Buneman, Cheney, Tan, & Vansummeren, 2008; Codd, 1970; Palmer, Zavalina, & Fenlon, 2010; Thibodeau, 2002). Finally, best practices in data collections curation must account for the fundamental need to migrate them over time. Where LAM collections are typically intended to last for generations, *digital* collections must necessarily rely on hardware that lasts years at best. Migration from one system to another is critical and inevitable given and the rate of obsolescence and decay in software, hardware, and storage media.

Though there is some guidance for database management and migration over time, much of it is focused on supporting large business-oriented platforms. For instance, Brodie and Stonebraker offer a guide through the general steps of database migration; however their work is targeted at business contexts with large homogenous data stores and dedicated information technology staff (Brodie & Stonebraker, 1995). Compelling work by Herrmann et al. proposes a database evolution language to support the migration of schemas over time, but further development would likely be needed for this approach to be feasible in many LAM contexts (Herrmann, Voigt, Rausch, Behrend, & Lehner, 2017). Extract-transform-load procedures (e.g. Henry, Hoon, Hwang, Lee, & DeVore, 2005; Vassiliadis, 2009) are similarly designed for large data warehouses and not the idiosyncratic and sometimes piecemeal systems of LAMs and small research labs. Additionally, these best practices don’t necessarily account for the on-going work of database curation over time. Maintaining a database for continual use as a research tool requires many hours of digitization, annotation, and updating, as well as provenance management (Buneman et al., 2006, 2008; Buneman, Müller, & Rusbridge, 2009). Many database systems are simply not designed to support these complex tasks, let alone over the course of multiple migrations.

Databases in NHMs

NHMs have been treating their collections catalogs as research datasets for decades; consequently, they are an excellent grounding case for this study. As early as the 1960s, NHMs began “computerizing” or “digitizing” (manually transcribing) their paper and card catalogs into databases, and by the 1970s started sharing records via information systems such as SELGEM and TAXIR, two information retrieval systems for taxonomic data (Hudson, Dutton, Reynolds, & Walden, 1971; Mello, 1975; Sarasan, Neuner, & Association of Systematics Collections, 1983). Several other community-driven data and Collection management platforms have been developed since then (notably Specify and Arctos, two open source NHM-specific Collection management systems), as well as numerous data sharing networks (for instance, MaNIS, the Mammal Networked Information System (Stein & Wieczorek, 2004); FishNet, a data sharing network for ichthyological collections (Vieglais, Wiley, Robins, & Peterson, 2000); and perhaps most well-known, GBIF (Robertson et al., 2014).

It is estimated that the grand “dataset” of all NHM collections contains anywhere from 1-2 *billion* specimen records; however, only a small fraction of these are currently accessible through data sharing platforms (Ariño, 2010). The vast majority of museum collections are not fully cataloged; and many of the catalogs that do exist still await digitization (Beaman & Cellinese, 2012). Thus, digital databases are often only partial representations of their collections – and, as Bowker notes, the natural world (Bowker, 2000). However, as Strasser argues, NHM collection databases “are not mere repositories; they are tools for producing knowledge" (Strasser 2011, pg. 63); the collections themselves are infrastructure to support scholarship, and as such have value beyond their completeness.

The impact of relational databases on collections data

NHMs are additionally informative for their use of relational databases for data collection management. Relational databases have been fixtures in many offices and research labs for decades (despite predictions of their impending obsolescence (e.g. Atzeni et al., 2013)), and they play an important, complex and often shifting role in information ecologies (Buneman et al., 2008; Hine, 2006). By relational databases, we mean those that use (or are built with software using) E. F. Codd’s relational data model. Codd’s approach to data representation was later formalized as the ANSI-SPARC three level architecture, which conceptualizes a data bank’s structure as: 1. A *physical* level where data are recorded on a storage medium; 2. A *logical* level, expressing relationships between data, and; 3. a *view* level, providing query access to stored data, but hiding the details of its organization from an end-user (Tsichritzis & Klug, 1978).

Codd intended this model to “[protect] users from having to know how data is organized” in a computer system (Codd, 1970), and argued that, “there is one consideration of absolutely paramount importance – and that is the convenience of the majority of users” (Codd, 1971). Relational architecture is thus predicated on a division of labor between database designers/administrators (who work at the logical level), and end-users (who work at the views level). Despite Codd’s best intentions, though, relational databases remain challenging to use. Part of the issue may lie in how “administrator” and “user” roles have blurred over time (see Dourish & Edwards, 2000; Jagadish et al., 2007; Li & Lochovsky, 1996; Olson, 2009; Voida, Harmon, & Al-Ani, 2011 for examples of the impact of blurred database roles). Where databases would have once been programmed by computer scientists into mainframe computers, they can now be run on a desktop, programmed by anyone with Microsoft Access or FileMakerPro. In NHMs, they have traditionally been designed and maintained by collection managers who typically have training in some branch of natural history, but not necessarily database administration. By studying how they have adapted (and adapted to) database technologies for their work, we may better understand how to support other similar LAM users.

APPROACH

Six collection managers (CMs) were initially recruited for this study based on their known use, management or design of natural history databases; an additional six were selected through snowball sampling. In total, twelve CMs at eight different NHMs were enrolled in this study. We sought representation across different types of institutions involved in natural history research (e.g. both universities and independent museums), and different departments within natural history museums. We did not seek a statistically representative sample of the databases or their users in NHMs; rather, we take a multi-case study approach and follow a replication logic, to identify common phenomena and patterns between cases (Yin, 2009).

We engaged our participants through three phases:

1. *Initial demographic survey and background research*: Each participant was asked to fill out a demographic survey, through which we gathered basic information about their background and training, as well as information about the number and nature of databases they manage through their work. We also asked participants for links to or copies of their databases.
2. *Initial interview*: Each participant was interviewed following the same semi-structured protocol, though some questions were tailored based on the participants’ survey response or based on our qualitative review of and databases or schemas they shared. We asked participants to describe this history of each of their databases. We asked them to describe three phases of each database’s evolution: 1) Its state at the time that they began working at the museum (as well as his or her knowledge of any prior versions); 2) The database’s present state, and; 3) Its future or anticipated state. By gathering data on these three different states, we could compare databases across time, departments, and institutions - seeking patterns in the way that content is normalized, migrated, and occasionally transferred between different database designs.
3. *Follow up interviews*: Between 12-15 months later, participants were contacted for follow up interviews, in which we received updates on their migration work. Eight CMs participated in follow-up interviews.

Data collection began in Spring 2014 and concluded in Winter 2016. Interviews lasted 45-75 minutes (60 minutes on average). Transcripts were summarized into short case reports of describing the history, migration and anticipated future state of each database. Transcripts were then coded for emergent themes related to database migration or evolution, and to identify CMs’ strategies for database curation; these were added to case reports. Common maintenance strategies and migration patterns were then derived from intercase comparison.

RESULTS

Our participants had anywhere from two to thirty years of experience working with databases. All but two were self-taught and learned their skills on the job through webinars, workshops, and textbooks. One CM said he, “bought a book on Access in 2010 and started on page one” (07-VertPaleo); another said she learned the basics of relational database construction by “googling for various manuals for FileMaker Pro” (01-InvertZoo) and studying them as she explored her files. For most, database migration was one of their first tasks when starting their jobs as CMs. In only 3 cases, participants were able to directly consult with their databases’ original creator; the rest had to reconstruct their databases’ history and structure through a quasi-forensic examinations of the database itself. This reverse engineering process entails extensive review of available documentation, careful examination of data structures and contents, and conversations with senior staff who may have been around during the database’s creation and fill in “tacit” details that lead to its creation.

Thirty-seven databases were described and managed by our participants, with a median of three per department. Database systems included Access, FileMaker Pro, MySQL Workbench, Arctos, Specify, KE EMu and Excel. We categorized databases into three types: Collection management (n=29), research (n=5) and auxiliary (n= ~6) (See tables 1-3). Below, we describe these databases according to their type more fully, and briefly discuss our participants’ database histories.

ID	Database contents	History
01-InvertZoo (n=3 total)	Mollusk collections (1)	Originally paper catalogs; digitized to Filemaker Pro; **migrated to Excel;** planned for migration to Arctos
	non-Mollusk invertebrate collections (1)	Originally paper catalogs; digitized to Filemaker Pro; **migrated to Excel;** planned for migration to Arctos
	Type specimens (specimens cited as exemplars of their species) (1)	Originally paper catalogs; digitized to Filemaker Pro; planned for migration to Arctos
01-VertZoo (n=17)	4 databases for each of 4 departments (birds, mammals, fish, reptiles) (16 total)	Each was transcribed from catalogs into FileMaker Pro; **migrated to Excel and OpenRefine for cleaning; migrated to Arctos;** keeping in Arctos for now
	Invertebrate paleontology collections database (1)	Originally created in Access of FileMakerPro; migrated to Specify 5; **migrated to Specify 6;** planned migration to Specify 7
02-VertPaleo (n=1)	Vertebrate Paleontology collections database (1)	Originally stored in a paper card catalog; likely directly digitized to Access database consisting of 2 unlinked tables; **migrated to Specify 6;** planned migration to Specify 7
03-Paleo Mineral (n=3)	Paleobotany database (1)	Originally stored in paper card catalog; digitized at some point; migrated to FileMakerPro; **migrated to Access; attempted migration to Specify; now each table of database stored as a separate spreadsheet;** planned migration to custom Microsoft SQL Server
	Mineral database (1)	Originally stored in paper card catalog**; transcribed to Excel**; planned migration to a custom Microsoft SQL Server database
04-Paleo (n=1)	Collections database (1)	Originally paper card catalog; digitized, possibly directly to Dbase; **migrated to Specify 5;** planned migration to Specify 6
05-InvertZoo (n=2)	Lot/collections database (1)	**Digitized from labels on specimen buckets in the Marine Biodiversity Lab directly into FileMakerPro**; currently maintained in FMP, but planned for migration to Specify 7
	"prospective database"- a database being "fielded" but not currently used" (1)	**Drafts of interfaces and schema maps for Specify 6 are now being migrated to Specify 7**
06-VertPaleo (n=1)	Vertebrate Paleontology Collections database (1)	Originally paper card catalog; digitized, possibly directly to Dbase; **migrated to Access**; planned migration to Specify 7
07-VertPaleo (n=2)	Collection management database (1)	Originally digitized from paper catalogs to Paradox (using punch cards!); migrated to Excel for some cleaning; **migrated and maintained in KE Emu;** no plans for migration
	Radiocarbon dates (1)	**Created in and maintained in Excel**; potentially to be migrated into KE Emu at some point.

Table 1. *Collection management databases. The ID column represents the institution and collection code and includes a total count of collections databases managed by the collection’s CM(s). The **Bolded** text in each database’s history indicates work undertaken by current collection manager (our study participant).*

Collection management databases: uncertain origins and varied legacy structures

Collection management databases are primarily used to support the management, curation and use of a physical specimen collection by organizing and preserving data about the specimens, though some have public-facing web interfaces that can be searched for research discovery. These databases have typically been migrated from paper catalogues and were first and foremost designed as a finding aid for the physical collections, though often the specimen records are published to aggregators such as GBIF and used as data in further analyses. In three cases (04-Paleo, 01-InvertZoo, and 05-InvertZoo) electronic collections databases have essentially replaced paper catalogues; in the rest they are maintained concurrently with paper catalogues, which are sometimes viewed as being more stable and long-lasting. Many of the collection management databases described by our participants had uncertain origins: each had been transcribed from paper records at some point in the past, but it was often unclear when and by whom.

Though we expected to find some instances of active database migration, we were surprised by how many of our participants were actively engaged in migration. At the time of our interviews, all but one collection database described by our participants was in the process of being migrated or was being prepared for a migration in the near future. An additional 2 CMs were in the midst of planning migrations of their *physical* collections. Participants broadly characterized database migration less as a single activity undertaken within a constrained time period, but instead, as an on-going (albeit often interrupted) aspect of their day-to-day work. They estimated that spent anywhere from 20-80% of their week on database work. Many expressed frustration that they did not have the time needed to complete their data-related activities. For instance, one CM (05-InvertZoo) described his two-year-long-effort to migrate an invertebrate zoology collections database as one of continual disruption:

> The Specify database [to which the legacy system would be migrated] was designed, but for lack of developer time, has not been fully fielded. I have been so interrupted for the past several years, that we have multiple times gotten to the 95% complete phase for introduction and migration into it, and then something comes up, we get distracted, and then six months later we try it again.

Other participants also described migrations that were *almost* completed, but eventually had to be abandoned for a variety of reasons. In two cases, CMs spent months considering platforms that turned out to be a poor fit for their specific collections (04-Paleo; 05-InvertZoo) and eventually had to start over from scratch.

Platformization of NHM collections databases: moves to Specify and Arctos

All but two of the collections databases described by our participants had been or were in the process of being migrated to Arctos or Specify, the two main community-developed databases being adopted by NHMs. Both are designed specifically for NHM collections and include pre-defined data schemas designed to be generally applicable to natural history collections. At the time of our interviews, there was a key difference between the two systems: Arctos is cloud-hosted, and publishes collections online to a public, aggregated database. On the other hand, Specify (at the time of our interviews) was locally hosted and did not publish or aggregate data for its users, but thereby allowed users some greater flexibility modifiability. (Since the time of our interviews, Specify has since developed a cloud-hosted service.)

Participants said that adopting these platforms made it easier to share or manage their data in a community-driven and vetted format; however, there were also some unexpected complications. Some users of Arctos found that they needed to extensively rework their legacy data schemas by hand to facilitate migration. Two users of Specify (03-PaleoMineral and 04-Paleo) had to “co-opt” fields – meaning, use a data field for other than its intended purpose – to map their legacy databases to the Specify data model. This was necessary because Specify either did not have a needed data field, or because Specify fields have technical constraints (numerical constraints, prohibiting certain characters) that conflicted with existing cataloging practices. Co-opting fields is common enough in the Specify community that the development team has addressed it in some of their documentation (“Importing External Data into Specify 6,” 2013).

While co-opting fields can solve cross-walking dilemmas in the short-term, it can create more problems in the long-term. For instance, the 04-Paleo collections database was migrated from dBase to Specify in the early 2000s, but the database now still retains evidence of its past structure. In the legacy database, each data field could only fit a maximum of 256 characters, so long locality descriptions had to be split between multiple fields (e.g. “stratigraphy 1,” “stratigraphy 2” and so on). These fields were never concatenated in subsequent migrations. Now, any query for stratigraphic data needs to be run over a range of fields, rather than just one. The CM described these traces as “ghosts” that haunt her in her day-to-day work: shadows of database structures past that hamper the habitability and functionality of her current system. The 01-InvertZoo CM reported similar hauntings from her predecessor’s co-opting of fields in Specify 5.

Auxiliary databases: spreadsheets as safe zones

CMs also maintained a range of *auxiliary databases* for specific tasks or data types that weren’t supported by their Collection management databases (Table 2). We include in this category systems that may not be typically considered databases but are used by our participants in a database-like fashion: Excel, OpenRefine, Google Sheets, and so on. Some of these are used for specific “work tasks” (04-Paleo), such as managing loans, printing labels, and organizing literature.

Participants reverted to spreadsheets for a number of reasons. First, the familiar interface simply made data entry easier, particularly for CMs relying on other staff or volunteers for data entry. Second, spreadsheets are materially separate from the collection database, and could therefore be used as a “safe” intermediary or staging area between raw data entry and the fidelity of the collection at large. This again makes them attractive for data entry; both 01-VertZoo and 07-VertPaleo used spreadsheets for student and volunteer data entry because they liked being able to review the data before adding it to the collection database. Finally, spreadsheets support kinds of data manipulations that simply are not possible in a "real" database: dragging and dropping values between fields; complex text string-editing functions; clustering and batch correcting entries; and so on. For instance, the CMs at 07-VertPaleo have recently begun several projects in which they loan out large numbers of microspecimens for identification and protein extraction. Using cloud-based Google Sheets, they share a copy of relevant collections records with their collaborators, who add data as it is generated. At the end of the project, the CMs re-import the newly improved records into the KE-EMu collection management database. This workflow is easy for everyone involved – and simply isn't possible with KE-EMu alone.

ID	Auxiliary database type	History	Purpose
01-Invert-Zoo	"Peripheral taxonomy database"	Originally paper; digitized to Filemaker Pro; **migrated to Excel;** planned for migration to Arctos	External storage of taxonomic hierarchies
03-Paleo Mineral	Research Libraries	Stored and **managed in EndNote**	Reference management, includes literature that cites specimens from the collection
	Web accessible collections database	**Created as Access database w/SQL backend**	web accessible front end for external researchers and visitors
04-Paleo	various "work tasks" databases	**Created as Access databases**; no plans to migrate	Label printing and loan management
07-VertPaleo	Field note database	Originally created in Access; **migrated to Excel; currently used for data entry in Excel;** will be merged w/KE Emu	data entry and temporary storage of data about uncatalogued specimens
	Various google spreadsheets used to collaborate with external researchers	**Created in Google Sheets; periodically migrated into KE Emu**	data entry and temporary storage of data about uncatalogued specimens

Table 2*. Auxiliary databases. The* ***Bolded*** *text again indicates work undertaken by our study participant.*

Spreadsheets-as-databases were also used as longer-term staging zones between database migrations. The 01-InvertZoo collection database became so denormalized and difficult to navigate in its legacy format that the CM exported the entire database into Excel for cleaning and storage until she was ready to migrate to Arctos. In a more extreme example, the 03-PaleoMineral database has been stored entirely in Excel since its planned migration to Specify failed. Each individual table of the database is being stored as a separate Excel file until their Microsoft SQL server database can be finished. Its CM is now managing and updating each table, including primary keys *by hand*.

Research databases

Five "research databases" were described by our participants (Table 3). We define a research database as one that has been created to support scientific projects and answer specific research questions. They are integrative in nature, drawing from many different collections and sources of data, and are meant to facilitate data analysis in addition to storage and retrieval. They may include some of the same content as a collections database (specimen records, locality descriptions), but are maintained as separate, custom-built databases, often in MySQL or Access. All five of the research databases described by our participants were developed to support specific scientific goals and were at least initially grant-funded.

ID	Database name	History	Purpose
01-Invert-Zoo	MapStedi	Originally created in FileMakerPro**, migrated to Excel**; future status is unclear	a mapping "toolkit for the southern and central Rockies and adjacent plains."
02-VertPaleo	MioMap	Both originally created in MySQL; **current CM maintains in MySQL;** planned merge into Neotoma	Paleoecology database of mammals and their habitats ~5-30 million years ago
	FaunMap		Paleoecology database of mammals and their habitats ~40,000 to 500 years ago.
04-Paleo	Neogene Marine Biota of the Tropical Americas (NMITA)	Originally created and **currently maintained in Oracle**; no plans to migrate	public database of photos and taxonomic information about the Neogene Marine Biota of the tropical Americas
05-Invert-Zoo	Systematics of Decapoda	Originally created in and **currently maintained in FileMakerPro w/ web-accessible MySQL mirror**; no plans to migrate	synthesizes the taxonomic literature of decapods (ten-footed crustaceans)
08-Ento	United Chalcidodea Database (UCD)	Created and **maintained in Paradox**; exported to a SQL "dump"; planned for migration to TaxonWorks	synthesizes the taxonomic literature of chalcid (tiny, parasitic, hyperdiverse) wasps

Table 3. *Research databases. The* ***Bolded*** *text again indicates work undertaken our study participant.*

Because research databases are built for specific projects, their underlying structure is correspondingly idiosyncratic. Several feature instances of deliberate denormalization: structures or attributes that went against what might ordinarily be considered best practices in database construction for the sake of the research workflow, or the broader research goals. The MioMap, FaunMap and Systematics of Decopoda databases show the clearest examples of this. In the case of MioMap and FaunMap, the two systems have the same relational structure and can even be queried through the same web interface – but are still maintained as separate databases. The CM (02-VertPaleo) explains that this siloing is for both scientific and social reasons. Because Miocene data is of a slightly different temporal granularity, keeping it in a separate database makes it easier to maintain, and also, “makes it easier for the Miocene workers to focus on the Miocene stuff." Additionally, he says that, “a lot of it has to do with feelings of ownership and accessibility. There’s a long history in paleontology of groups feeling like they don’t want to have to share control of the data with other groups.” Keeping the two databases separate helps groups feel like they have control over their own materials.

In the case of the Systematics of Decopoda (“ten-footed” crustaceans like lobsters and crabs) (SoD) database, strategic denormalization takes the form of the strategic preservation of historical typos. The SoD database is a taxonomic database: a carefully curated collection of species descriptions in published literature. These may superficially seem like a simple bibliographic resource, but 05-InvertZoo explains,

> It’s important to remember that for taxonomists, the bibliographic information are data. It’s not just the description of where to find the reference. It’s actually data that represent the publication date, the publication information for a particular taxon concept, and they are very, very careful about that information.

The SoD curator and his colleagues have been careful not to “improve” the data quality as records were input, because “the systematists involved in the project had a very strong insistence that the bibliographic data exactly reflect the publication as it existed... This was not an attempt to rectify the literature, it was an attempt to reflect the literature as it existed.” Any typos or misspellings in legacy citations are painstakingly preserved through two provenance tracking mechanisms: an annotation field where data enterers must justify any changes, as well as a version control system that archives edits. The database is now fairly widely used by decapod researchers (one hundred use it regularly, and a few thousand more casually). Our participant credits the aggressive provenance-tracking and annotation with making it usable by other scholars:

> [When] we reflected with the data that we made available, the change history of the records, all of a sudden, it gave a huge amount of credibility because then they can see what level of trust they’re willing to put on any individual record.... And we’ve had feedback from the community that having that, exposing that kind of metadata about the metadata – the metadata metadata – is why they will very often tend to use our listing, rather than any other authoritative source where they are not able to check... the audit trail of the information.

DISCUSSION

We noted several trends in the maintenance and migration of our participants’ databases. We summarize these below, and additionally identify points for future tool development or further research.

Maintenance through adaptation: strategies to make the database work

Though participants drew on a range of techniques to manage their databases, the following were particularly pervasive:

- *Strategic denormalization.* Several participants deliberately denormalized databases to facilitate data entry, or otherwise keep similarly structured data separate for different user groups. In doing so, CMs are essentially adapting the mechanics of relational databases to support security, privacy or usability needs. We note that Jagadish et al., similarly found that denormalization was sometimes necessary to make databases more usable (Jagadish et al., 2007).
- *Co-opting fields*. Users of databases with pre-determined data models and data entry interfaces (notably, older versions of Specify) sometimes co-opted fields for other-than-their-intended purpose. By doing so they are able to alter the database for their particular context without having to make changes at the schema level.

Both of these strategies make databases fit for immediate and idiosyncratic use but can cause problems over time. Denormalized tables are more challenging to consistently update and will inevitably require extensive cleaning before being made fit-for-use; and co-opting fields can make the schema mapping process of a database migration incredibly challenging. Both strategies risk leaving a database “haunted” (as our participant described it) by traces of prior use and users.

These maintenance strategies also double as database migration *avoidance* strategies. The desire to keep data in the same system is incredibly strong; the creative steps CMs take to keep their databases running despite their shortcomings is a testament to this inertial force. Our participants’ databases were typically migrated only when legacy platforms became absolutely unsustainable (e.g. no longer capable of running on modern computers, or no longer supported by other IT staff), or because their creators/former custodians retired. To borrow from Kuhn: database paradigms don’t really change, but rather, their advocates simply retire! This as a potentially self-perpetuating cycle. If database migrations are only reactionary (rather than proactive), they will likely be all the more painful and challenging; and if database migrations are always painful and challenging, then they are more likely to be delayed until absolutely necessary. There is a clear need, then, to further support proactive, incremental data collection migrations over time. Adaptation of existing best practices for incremental database migration (Brodie & Stonebraker, 1995) and recent work on database evolution languages (Herrmann et al., 2017) may facilitate this kind of long-term incremental migration.

Migration patterns: strategies to facilitate change

We noted several common themes in our participants' database migration strategies:

- *Reworking relationships*. In several cases, CMs had to transform legacy data schemas to fit new databases, reworking relationships between records and specimens, or between different tables in information models. This work was largely done "by hand," through manual manipulation of columns and rows within tables. We note that while there are tools for schema migration, none of our participants used them; further work is needed to explore why.
- *Community consultation, and duplication with adaptation*. We couple these themes because they are closely linked. All of our participants said that they spoke extensively with their fellow collection managers before beginning a database migration or selecting a new database system; and in some cases, they modelled data schemas or interfaces off of one another's systems. Even beyond the use of community-developed databases, community support and advocacy plays an important role in the selection, design and maintenance of collections databases.
- *Reliance on spreadsheet software*. Where others have noted the impact of the "psychological heritage of print" on users' wayfinding in databases (Kerr, 1990), here we note the psychological heritage of the spreadsheet for data entry and manipulation. Numerous participants used spreadsheet software as a place to temporarily store data (though we note that some took a fairly long view of temporary – months to even years at a time) between migrations. Others use them as a *lingua franca* to ease collaboration between near and distant project partners.

Platformization, infrastructuralization, and technological appropriation

One common theme – the move to community-developed databases – is notable enough to merit a more extended discussion. All but two of the collections databases described by our participants had been or were in the process of being migrated to Arctos or Specify; an additional three research databases (MioMap, FaunMap and UCD) were being integrated with or into larger community-developed infrastructures. Given our participants' descriptions of broader trends and our own prior experience in NHMs, we believe that our participants' migrations are representative of a discipline-wide movement: NHM data managers are moving away from in-house management of data infrastructure and moving toward NHM-specific platforms such as Arctos, Specify, KE-Emu, and Symbiota, among others.

We were initially tempted to characterize this trend as the "platformization" of NHM collections infrastructures – an adoption of middleware that imposes a top-down standard (specifically, a standardized data schema), around which users develop and maintain complementary components (in this case, auxiliary databases and query interfaces) (Plantin, Lagoze, Edwards, & Sandvig, 2016). However, the move toward primarily *community-developed*, open source and non-profit data management platforms complicates this account. For decades, NHMs have been using a technology that was not directly built for them: relational database software, which was designed for business use, and not for the iterative work of scientific data curation and management. The creative ways in which NHM workers have maintained their databases through adaptation can be viewed as a kind of technological appropriation by a (technologically) marginalized community (Eglash, 2004). As Eglash argues, technological appropriation occurs when groups with low social power "re-conceptualize ideas and artifacts" and "become producers" in their own right. While NHM staff, like all scientists (and particularly, Western scientists), certainly occupy positions of power in many ways, they have historically had relatively little sway over the computational tools at their disposal. But with the advent of community-developed database management systems, the technologically underserved NHM community has become the producers of its own database technology, information standards, and data publishing mechanisms.

Further work is needed to fully explore the nuances of how appropriation of database technology has impacted the NHM community, as well as other LAMs. Eglash notes that technological appropriation by those with low social power has potential to a community "move toward strong democracy"; however, there is also a risk that existing undemocratic hierarchies are simple replicated in the new NHM environment. For instance, there are already hints at a replication of the divide between users and producers within NHMs (e.g. users vs producers of Arctos), as well as a replication of the environment that necessitated appropriation in the first place. By co-opting fields, Specify users have had to make their own adaptations of an already adapted technology. This implies that users are still finding a need to innovate on their technology. Additionally, we note that not all users have found a solution to their data management woes; some, like the CM at 03-PaleoMineral, have found that the "community" developed products simply don't work for her edge case.

CONCLUSION & FUTURE WORK

In this study we have described database maintenance and migration work by collection managers at NHMs. We have identified some common strategies they deploy in adapting systems for continued use and maintenance, and in migrating their databases over time. We have additionally described how this community's move toward community-developed systems may reflect a kind of technological appropriation – albeit one that may risk reifying existing divides between users and developers. We believe these findings have implications not just to those working in NHMs or with NHM data, but also to others interested in the long-term maintenance of scholarly collections or datasets.

In our future work, we plan to look beyond NHMs to research collection databases used in other fields, and thereby expand on our understanding of data collection maintenance strategies and migration patterns. Though our project is still in early stages, we have identified several preliminary implications for practitioners, which we would like to explore further. For instance, we find that planning for truly long-term – on the scale of decades – digital collection curation requires the consideration of topics not always discussed in database design, such as the need to plan for retirement both of people and of information technologies. We also find that data collections migration is less of a one-time event and more a process that requires on-going consideration and preparation; we'd like to explore ways of making this interim migration work more intentional amidst information professionals' otherwise busy jobs. Finally, we find that data collection migration involves innovation with the resources at hand, and occasionally appropriating existing tools in unexpected ways; we hope to explore what tools and workplace conditions best foster this innovation and identify ways to support information professionals in sharing their *ad hoc* strategies.

ACKNOLWEDGMENTS

Enormous thanks to our study participants for sharing their time and insights for this project; and thanks to the UMSI Information and Archival Science reading group for feedback on early drafts of this paper.

REFERENCES

Ariño, A. H. (2010). Approaches to estimating the universe of natural history collections data. Biodiversity Informatics, 7(2). https://doi.org/10.17161/bi.v7i2.3991

Atzeni, P., Jensen, C. S., Orsi, G., Ram, S., Tanca, L., & Torlone, R. (2013). The relational model is dead, SQL is dead, and I don't feel so good myself. ACM SIGMOD Record, 42(1), 64. https://doi.org/10.1145/2503792.2503808

Beaman, R., & Cellinese, N. (2012). Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys, 209, 7–17. https://doi.org/10.3897/zookeys.209.3313

Brodie, M. L., & Stonebraker, M. (1995). Migrating legacy systems: gateways, interfaces & the incremental approach. San Francisco, Calif. : [S.l.]: Morgan Kaufmann Publishers ; IT/Information Technology.

Buneman, P., Chapman, A., & Cheney, J. (2006). Provenance management in curated databases (p. 539). ACM Press. https://doi.org/10.1145/1142473.1142534

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

Buneman, P., Müller, H., & Rusbridge, C. (2009). Curating the CIA World Factbook. International Journal of Digital Curation, 4(3), 29–43. https://doi.org/10.2218/ijdc.v4i3.126

Chan, S. (2014, November 7). The API at the center of the museum. Retrieved June 8, 2018, from https://labs.cooperhewitt.org/2014/the-api-at-the-center-of-the-museum/

Codd, E. F. (1970). A relational model of data for large shared data banks. Communications of the ACM, 13(6), 377–387. https://doi.org/10.1145/362384.362685

Codd, E. F. (1971). Normalized data base structure: a brief tutorial (p. 1). ACM Press. https://doi.org/10.1145/1734714.1734716

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 368(1926), 4023–38. https://doi.org/10.1098/rsta.2010.0165

Dourish, P., & Edwards, W. K. (2000). A Tale of Two Toolkits: Relating Infrastructure and Use in Flexible CSCW Toolkits. Computer Supported Cooperative Work (CSCW), 9(1), 33–51. https://doi.org/10.1023/A:1008709725729

DPLA API Codex. (n.d.). Retrieved June 8, 2018, from https://pro.dp.la/developers/api-codex

Eglash, R. (Ed.). (2004). Appropriating technology: An introduction. In Appropriating technology: vernacular science and social power (pp. vii–xxi). Minneapolis: University of Minnesota Press.

Henry, S., Hoon, S., Hwang, M., Lee, D., & DeVore, M. D. (2005). Engineering trade study: extract, transform, load tools for data migration. In 2005 IEEE Design Symposium, Systems and Information Engineering (pp. 1–8). https://doi.org/10.1109/SIEDS.2005.193231

Herrmann, K., Voigt, H., Rausch, J., Behrend, A., & Lehner, W. (2017). Robust and simple database evolution. Information Systems Frontiers, 20, 45–61. https://doi.org/10.1007/s10796-016-9730-2

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

Hudson, L. W., Dutton, R. D., Reynolds, M. M., & Walden, W. E. (1971). TAXIR-A biologically oriented information retrieval system as an aid to plant introduction. Economic Botany, 25(4), 401–406. https://doi.org/10.1007/BF02985207

Importing External Data into Specify 6. (2013, August 17). Specify Software Proejct. Retrieved from http://www.sustain.specifysoftware.org/wp-content/uploads/2017/03/Importing-External-Data-into-Specify-6.pdf

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

Kerr, S. T. (1990). Wayfinding in an electronic database: The relative importance of navigational cues vs. mental models. Information Processing & Management, 26(4), 511–523. https://doi.org/10.1016/0306-4573(90)90071-9

Li, Q., & Lochovsky, F. H. (1996). Advanced database support facilities for CSCW systems. Journal of Organizational Computing and Electronic Commerce, 6(2), 191–210. https://doi.org/10.1080/10919399609540276

Mello, J. F. (1975). The Use of the Selgem System in Support of Systematics. In Computers in Botanical Collections (pp. 125–138). Springer, Boston, MA. https://doi.org/10.1007/978-1-4684-2157-6_17

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

National Science and Technology Council, & Scientific collections: mission-critical infrastructure for federal science agencies: a report of the Interagency Working Group on Scientific Collections (Eds.). (2009). Scientific collections: mission-critical infrastructure for federal science agencies: a report of the Interagency Working Group on Scientific Collections. Washington DC: Office of Science and Technology Policy.

Olson, J. E. (2009). Database archiving: how to keep lots of data for a very long time. San Francisco, Calif. : Oxford: Morgan Kaufmann ; Elsevier Science [distributor].

Padilla, T. G. (2018). Collections as data: Implications for enclosure. College & Research Libraries News, 79(6). Retrieved from https://crln.acrl.org/index.php/crlnews/article/view/17003/18751

Palmer, C. L., Zavalina, O. L., & Fenlon, K. (2010). Beyond size and search: Building contextual mass in digital aggregations for scholarly use. Proceedings of the American Society for Information Science and Technology, 47(1), 1–10. https://doi.org/10.1002/meet.14504701213

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2016). Infrastructure studies meet platform studies in the age of Google and Facebook. New Media & Society, 1461444816661553. https://doi.org/10.1177/1461444816661553

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., … Desmet, P. (2014). The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS ONE, 9(8), e102623. https://doi.org/10.1371/journal.pone.0102623

Sarasan, L., Neuner, A. M., & Association of Systematics Collections (Eds.). (1983). Museum collections and computers: report of an ASC survey. Lawrence, Kan., U.S.A: Association of Systematics Collections.

Stein, B. R., & Wieczorek, J. R. (2004). Mammals of the World: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics, 1(0). https://doi.org/10.17161/bi.v1i0.7

Tate. (2018). collection: Tate Collection metadata. Python, Tate Modern. Retrieved from https://github.com/tategallery/collection (Original work published 2013)

Thibodeau, K. (2002). Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years (The State of Digital preservation: An International Perspective). Washington, D.C.: CLIR and the Library of Congress. Retrieved from https://www.clir.org/pubs/reports/pub107/thibodeau/

Tsichritzis, D., & Klug, A. (1978). The ANSI/X3/SPARC DBMS framework report of the study group on database management systems. Information Systems, 3(3), 173–191. https://doi.org/10.1016/0306-4379(78)90001-7

Vassiliadis, P. (2009). A Survey of Extract–Transform–Load Technology: International Journal of Data Warehousing and Mining, 5(3), 1–27. https://doi.org/10.4018/jdwm.2009070101

Vieglais, D., Wiley, E. O., Robins, R., & Peterson, T. (2000). Harnessing Museum Resources for the Census of Marine Life: The FISHNET Project. Oceanography, 13(3), 10–13. https://doi.org/10.5670/oceanog.2000.02

Voida, A., Harmon, E., & Al-Ani, B. (2011). Homebrew databases: complexities of everyday information management in nonprofit organizations (p. 915). ACM Press. https://doi.org/10.1145/1978942.1979078

Wickett, K. M. (in press). A logic-based framework for collection/item metadata relationships. Journal of Documentation.

Wickett, K. M., Renear, A. H., & Urban, R. J. (2010). Rule categories for collection/item metadata relationships. Proceedings of the American Society for Information Science and Technology, 47(1), 1–10. https://doi.org/10.1002/meet.14504701218

Yin, R. K. (2009). Case study research: Design and methods. (4th ed.). SAGE Publications Ltd.

Zavalina, O. L., Palmer, C. L., Jackson, A. S., & Han, M.-J. (2009). Evaluating Descriptive Richness in Collection-Level Metadata. Journal of Library Metadata, 8(4), 263–292. https://doi.org/10.1080/19386380802627109