Logic-Based Schema Alignment for Natural History Museum Databases†

Andrea Thomer*, Yi-Yun Cheng**, Jodi Schneider**,
Michael Twidale**, Bertram Ludäscher**

*University of Michigan, School of Information, 105 S. State St., Ann Arbor, MI 48109,
<athomer@umich.edu>
**University of Illinois at Urbana-Champaign, School of Information Sciences,
501 E. Daniel St Champaign, IL 61820,
<[yiyunyc2, jodi, twidale, ludaesch]@illinois.edu>

Andrea Thomer is an assistant professor of digital curation. She earned her doctorate at the School of Information Sciences at the University of Illinois at Urbana-Champaign in 2017. She conducts research in the areas of digital curation, natural history museum informatics, information organization, and information system usability. She is particularly interested in the long-term usability of digital collections and their infrastructures. Prior to her work in information science, she was an excavator at the La Brea Tar Pits, a rich fossil site in Los Angeles. She continues to draw on her experience in paleontology, evolutionary biology, and museums.

Yi-Yun Cheng is a PhD student at the School of Information Sciences of the University of Illinois at Urbana Champaign. She received her M.A. degree from the Department of Library and Information Science at National Taiwan University in 2015. Her previous work is related to the semantic web and ontology mapping; now she is working on taxonomy alignment projects. Her main research interest is the interoperability among ontologies, taxonomies, or other types of knowledge organization systems.

Jodi Schneider is an assistant professor in the School of Information Sciences at the University of Illinois at Urbana-Champaign. She studies scholarly communication and social media through the lens of arguments, evidence, and persuasion. She is developing linked data (ontologies, metadata, semantic web) approaches to manage scientific evidence. Schneider holds degrees in informatics (PhD, National University of Ireland, Galway), library and information science (MS, University of Illinois), mathematics (MA, University of Texas-Austin), and liberal arts (BA, Great Books, St. John's College). She worked in academic libraries and bookstores for six years.

Michael Twidale is a professor at the School of Information Sciences, University of Illinois at Urbana-Champaign. His research interests include computer supported cooperative work, computer supported collaborative learning, and human computer interaction. Current projects include studies of informal social learning of technology, metrics for open access, sociotechnical systems design, collaborative information retrieval, computational metacognition, agile research methods, and long term scientific database management. His approach involves the use of interdisciplinary techniques to develop high speed, low cost methods to better understand the difficulties people have with existing computer applications and so to design more effective systems.

Bertram Ludäscher is a professor at the School of Information Sciences at the University of Illinois, Urbana-Champaign, directs the Center for Informatics Research in Science and Scholarship, is a faculty affiliate with the National Center for Supercomputing Applications (NCSA) and the Department of Computer Science. His research interests range from scientific data and workflow management, to knowledge representation and reasoning. He received his M.S. in computer science from the University of Karlsruhe (now: KIT), and his PhD from the University of Freiburg, both in Germany.

Andrea Thomer, Yi-Yun Cheng, Jodi Schneider, Michael Twidale and Bertram Ludäscher. 2017. "Logic-based Schema Alignment for Natural History Museum Databases. *Knowledge Organization* 44, no. 7: 545-558. 48 references.

Abstract: In natural history museums, knowledge organization systems have gradually been migrated from paper-based catalog ledgers to electronic databases; these databases in turn must be migrated from one platform or software version to another. These migrations are by

no means straightforward, particularly when one data schema must be mapped to another—or, when a database has been used in other-than-its-intended manner. There are few tools or methods available to support the necessary work of comparing divergent data schemas. Here we present a proof-of-concept in which we compare two versions of a subset of the Specify 6 data model using Euler/X, a logic-based reasoning tool. Specify 6 is a popular natural history museum database system whose data model has undergone several changes over its lifespan. We use Euler/X to produce visualizations (called "possible worlds") of the different ways that two versions of this data model might be mapped to one another. This proof-of-concept lays groundwork for further approaches that could aid data curators in database migration and maintenance work. It also contributes to research on the unique challenges to knowledge organization within natural history museums, and on the applicability of logic-based approaches to database schema migration or crosswalking.

Received: 28 June 2017; Revised 26 September 2017; Accepted: 27 September 2017

Keywords: database schema, database migration, Euler/X, possible worlds

† Thanks to Tim Cole & Jerome McDonough for helpful discussions about metadata crosswalks, and to the DALS research group at the University of Michigan School of Information for feedback on this manuscript. This work was supported in part by NSF DBI-1643002.

1.0 Introduction

In natural history museums (NHMs), collections data often have longer lifespans than the knowledge organization systems (KOSs) used to make them accessible. Consequently, migration from one KOS to another is periodically necessary. When NHM KOSs were strictly made of paper, ink, and the arrangement of shelves and drawers—all relatively stable information storage formats—KOS migration happened perhaps once in a generation. However, following the move to primarily electronic collections databases beginning in the 1970s, NHMs now must migrate their entire catalogs as often as hardware and software updates dictate: every few years, rather than every few decades. Modern NHM KOS management consequently entails the frequent assessment, curation and migration of sometimes-complex relational database schema.

Migrating and managing data schemas over time is by no means straightforward; further, migrating data from one schema to another can result in unexpected information loss or alteration. For instance, if a NHM record is published from an idiosyncratic local schema to a public database such as the Global Biodiversity Information Facility (GBIF), the record may need to be crosswalked in a way that could risk altering its elements' meaning (see Thomer et al. 2012 for a brief discussion of this issue specific to NHMs; see also St Pierre and LaPlante 1998 for a general overview of issues related to crosswalking). Similarly, migrating legacy databases to newer, "off-the-shelf" systems that come with predetermined schema, i.e. NHM-specific databases such as Specify (http://www.sustain.specifysoftware.org/), Arctos (https://arctos.database.museum/), and KE Emu (https://emu.kesoftware.com/), can require unique workarounds to make legacy and/or locally-important data "fit" into the new structure. NHM collections managers have reported needing to "co-opt" fields within "off-the-shelf" databases for other-than-their-intended purpose, thereby effectively altering the prescribed data model to suit their local needs.

The effects of such changes to a schema, or of aberrant use of a schema, are subtle and often not immediately apparent. Whereas "physical" migrations from one organizational scheme to another (such as changes to a shelving system or cataloging style) can be seen by the naked eye, the impact of migration from one database schema to another typically requires logical analysis to be truly understood. Few tools exist for this work. Further research is needed to support the task of database migration, particularly for memory institution staff such as collections managers and curators who are certainly experts in their fields but not necessarily experts in database development. Additionally, further research is needed to support the development of tools that might help curators understand the subtle impact of idiosyncratic, aberrant, or otherwise unconventional database use and database migration.

In this paper, we address the intertwined issues of comparing an old and new version of the "same" schema, and understanding the impact of aberrant use of a field within a schema (such as the "co-opting" behavior described above) on database migration. We explore the utility of a taxonomy alignment tool, Euler/X, in revealing alignments and possible conflicts between two museum data schemas. Euler/X is a logic-based tool that employs a particular formalism called Region Connection Calculus (RCC-5) to compare and reconcile two or more taxonomies. RCC-5 calculates the five possible relationships between nodes of a taxonomy: congruence (c1=c2), inclusion (c1>c2), inverse inclusion (c1<c2), overlap (c1 o c2) and disjointness (c1 ! c2). For any two taxonomies, the relationships between their nodes can be determined by a domain expert or generated by the tool. Then, given two taxonomies along with their relationships, Euler/X tool can

create a "combined" or "merged" result taxonomy that reconciles the different perspectives represented by the input taxonomies. In this way, Euler/X can be used to compare and merge an "old" and a "new" taxonomy, or multiple overlapping taxonomies. Euler/X was developed specifically for reconciling multiple taxonomic "perspectives"—in other words, for logically "sorting things out" (with apologies to Bowker and Star (1999)). The efficacy of the Euler/X approach has been previously demonstrated through analysis of how botanists' classifications changed over time, in a use case involving alignments of eleven botanical classifications spanning one hundred twenty-six years (Franz, Pier, et al. 2016).

Here we use Euler/X to compare two versions of a subset of the Specify database schema. Specify is a popular NHM database system developed and maintained by the University of Kansas Biodiversity Institute. Specify 6's underlying database schema has undergone upwards of nine updates since 2008 ("Documentation" 2017). Using Euler/X, we compare and reveal differences or conflicts between a subset of Specify's original schema ("Specify 6 Schema" 2009) and its current version ("Specify DB Schema 2.3" 2016).

Our proof-of-concept analysis produces visualizations of the five "possible worlds" that result when trying to merge (reconcile) the two versions of the Specify 6 schema. A "possible world" is a potential solution to the taxonomy alignment problem; it shows a way in which two schemas "might" be mapped to one another. They might be thought of as parallel universes that represent all merged solutions from the consistent joint input conditions (Cheng et al. 2017). In Euler/X, different "possible worlds" correspond to different solutions to the underlying constraint satisfaction problem posed by a taxonomy alignment problem $T_1 + T_2 + A \rightarrow T_3$. We use this analysis to show how schema changes at an attribute level have impacts on the structure of KOS data schemas at higher levels, and discuss how this work reflects a need to support a plurality of KOS schemas. We additionally tie this to our prior work exploring "how databases learn" (Thomer and Twidale 2014) and discuss gaps in current database migration tools and migration documentation methods. This proof-of-concept lays groundwork for the development of tools that could be useful to data managers in their database migration work. It also contributes to an understanding of the unique challenges to knowledge organization within the NHM domain, as well as a discussion of the applicability of logic-based approaches to database schema migration or crosswalking.

2.0 Background

2.1 Natural history knowledge organization: from paper ledgers to electronic databases

Modern NHM KOSs are rooted in a long-standing tradition of natural history data collection and documentation practices. While methods of natural history data "analysis" have certainly become more computational, natural history modes of data collection and management are still remarkably similar to those used in the late nineteenth century. Researchers venture into the field alone or in small groups, collect specimens and other data, and record inventories of these specimens in their field books. These specimens are assigned field numbers corresponding to entries in the field inventories. The resulting inventories and field numbers are the basis for later specimen cataloging and labeling within the museum.

Understanding this historical context is important in any consideration of any modern NHM KOS. As Callery summarizes (1999, 85-6),

> The design and use of electronic information systems to provide access to natural history museum collections is influenced by existing traditions of organizing paper-based information about those collections …. In these museums the evidential value of the object itself is supplemented, not supplanted, by the documentary evidence of field notes, photographic and other visual records, formal accession information, and published works referring to that specific object.

In other words, NHM collections must first and foremost preserve and support access to physical specimens, and the KOSs used for this are rooted in diverse, distributed, paper-based systems. Because of the need to prioritize care of physical specimens, as well as the distributed nature of legacy paper-based KOSs in NHMs, many collections have had to digitize their catalogs in a piecemeal fashion when time and funding allowed (Berents, Hamer, and Chavan 2010). Consequently, modern digital NHM KOSs are often in a range of file formats and software platforms.

The schemas underlying these KOSs may also be structured in a manner idiosyncratic to the institution. There is no formal standard such as the library world's Resource Description and Access (RDA) framework for NHM cataloging; instead, a variety of best practices exist. One example of these best practices is the "Grinnell System" of recording field notes. Joseph Grinnell was a field biologist and the original director of the Berkeley Museum of Vertebrate Zoology (circa 1908). He developed a

method of field notetaking that dictates everything from what kind of ink to use ("The India ink and paper of permanent quality will mean that our notes will be accessible 200 years from now" (Grinnell 1958, 8)) to how and where one should record the date, time and place on each page. Grinnell taught this method to his colleagues and students at the Museum of Vertebrate Zoology, and it eventually became broadly adopted by field biologists and naturalists in other regions as well. A Grinnellean field notebook's structured "catalog" section, with prescribed fields and formats for the date, location, catalog number, species, sex, breeding status, and morphological measurements of a specimen might be viewed as an ancestor of the modern NHM database (Perrine and Patton 2011). Despite the structure offered by recommendations such as Grinnell's, however, there is still often necessary variation in different researchers' cataloging methods. Different institutions and domains of study have different needs of their data and must shape their practices accordingly (Bowker 2000).

These institutional- and domain-based idiosyncrasies were not necessarily problematic when NHMs first began creating databases for "local" access in the 1960s; however, in the 1980s, the move toward community-based data publishing infrastructures motivated the development of shared data standards. At this time, NHMs began federating and aggregating their collections online through platforms such as the Mammal Networked Information System, HerpNet, FishNet, and VertNet (Callery 1999). Organizations such as the Taxonomic Database Working Group and the Association of Systematics Collections formed to develop data models and standards such as the Darwin Core and Access to Biodiversity Collections Data standards and the ASC Information Model for Biological Collections (ASC 1993; Wieczorek et al. 2012; Berendsohn et al. 1999). Eventually, many museums began migrating their collections databases to community-developed "off-the-shelf" systems such as Specify and Arctos, which were designed to natively support data publishing.

These "off-the-shelf" databases all come with predetermined data schemas, relieving NHM collections staff of the need to create their own databases from scratch. However, this relief comes at a cost: legacy databases must be migrated or crosswalked to a new standardized schema. Alternately, collections staff must find ways of creating unconventional workarounds to fit idiosyncratic legacy data into standardized formats. One such workaround is to "co-opt" fields within the database for other-than-their-intended purpose (discussed in Brenskelle 2015). For instance, if a collection manager needs to record, say, the wingspan of a bird specimen, but there is not a predetermined field for wingspan, she might choose to use a field she doesn't otherwise need (perhaps, "radiocarbon date"). Co-opting fields can solve database migration problems in the short term, but can have difficult-to-predict consequences when the schema is changed by developers in the normal course of database updates. In particular, when the underlying database structure, or schema, changes in one of these "off-the-shelf" databases, any such local customizations will break. Thus, there are intertwined issues of aberrant database use and schema evolution at play in these KOSs over time.

2.2 Schema evolution and crosswalks

The need to understand how KOSs adapt to changes in knowledge, particularly over time, has been identified as an important question for research in knowledge organization (Gnoli 2008; Lauruhn and Groth 2016; ; Scharnhorst et al. 2016; Tennis 2012 and 2016). The problem of schema evolution is not new, and is not unique to NHMs (see Roddick 1992; see also Gao and Zaniolo 2012b; Brahmia et al. 2015a and b; Galante et al. 2005; Gao and Zaniolo 2012a). Schemas can evolve for a variety of reasons, including but not limited to:

- Changes in the purposes of data collection and associated scientific priorities,
- Changing, often more systematic, work practices that require greater precision or different data acquisition technologies;
- Evolving disciplinary, national, and international standards;
- A desire to work towards a greater harmonization and ultimately integration with other similar datasets for greater interoperability; and
- Changes in the software and hardware used for KO.

In the NHM context, schema evolution can involve:

- Changes in data collection and documentation practices: what is collected, how it is recorded, and the level of detail;
- Changes in how the data is represented;
- The addition of new fields to record additional information;
- Splitting fields, to record data in a more structured manner;
- Aggregating fields;
- Deleting fields; and
- Moving fields into different tables.

Over time there may be a trend towards collecting more data and in a more systematic way with greater use of controlled vocabularies and more fine-grained structure

through the creation of database subfields. Keeping track of these changes is challenging but important. Creating a crosswalk between different generations of schemas facilitates migration, and can also reveal unintended or unanticipated ambiguities between the old and new schemas.

2.3 Prior work on database migration and crosswalking.

The challenges of crosswalking data standards or models (that is, creating a specification to map one standard to another) have received considerable attention from the library and information science community. Consistent, harmonized metadata aggregated from multiple sources is often needed to support information retrieval in information systems such as union catalogs or data aggregators; harmonization may start by mapping between different metadata standards. Creating and maintaining metadata crosswalks is challenging but can reduce the cost of creating metadata while enabling interoperability (St. Pierre and LaPlant 1998).

A range of crosswalking resources are used in practice. Hand-curated crosswalks by single institutions have been shared in tabular formats (for instance, those created by the Getty Research Institute (Harping 2014)). Computable crosswalks and tools built on crosswalks also exist. For example the RDF ontology developed by the JISC Vocabulary Mapping Framework can be queried for the closest match between terms; it takes a hub and spoke approach, mapping each vocabulary to an extensible and semantically-rich central "hub" data model (JISC 2009). OCLC maintains a crosswalk web service that can translate from one metadata record standard, structure, and encoding to another ("Metadata Schema Transformation Services" 2014). Translations between XML-based metadata formats are sometimes implemented using XSLT stylesheets (e.g., "Conversions: Metadata Object Description Schema: MODS" 2017).

Regardless of the approach, preserving meaning is a key challenge of metadata crosswalking and database migration. Ambiguous or implicit semantics can cause problems when moving data from one schema to another. Correct treatment of a resource often depends on knowledge that is incompletely or imprecisely represented. For example, sometimes a record conflates multiple items—e.g., an image, the file that encodes it, the metadata description, and the software that stores the metadata description in a way that presents no problem to humans but which computers cannot interpret. Likewise (Dubin et al. 2009, 599), "crucial contextual data may exist only as natural language annotations or as unstructured information in the content of metadata fields."

The complexity of crosswalk development should not be underestimated. As Zeng and Chan (2006) note:

> The reality is that crosswalks constructed based on the real data conversion might be very different from those based on metadata specifications. Additional instructions and detailed explanations need to be provided for different situations. Unfortunately, most crosswalks are focused only on mappings based on metadata specifications, not on real data conversion results.

Lack of organizational memory can complicate crosswalking projects. Khoo and Hall (2010) describe challenges in crosswalking two digital libraries to the Dublin Core standard. In their work, they found multiple legacy databases that had not previously been migrated to the main library catalogs. Customized metadata fields were used but documentation for them was not available; this led to extensive discussions before ruling some data irrelevant to users. Many idiosyncrasies in the catalog data, especially local usage and changes in metadata practices over time, were not found until after the project was underway. Such idiosyncrasies may include elaborating existing categories, creating new subcategories, and adding higher order categories (Trigg, Blomberg, and Suchman 2002).

Euler/X cannot automate these complexities away, but we believe that it can be useful in highlighting the cause and exact nature of certain complexities. For example, the multiplicities of possible worlds (i.e., the different solutions to a schema alignment problem) that Euler/X highlights can expose inherent ambiguities in the given problem. Conversely, if no possible world exists, this means that not all input articulations A can be simultaneously satisfied. In other words, there are logical conflicts (contradictions) in A, even though different name spaces (here: terminologies in form of input taxonomies/schemas T_1 and T_2) are used. Often, these are exactly the same problems that will arise in multiple contradictory interpretations around data entry and data analysis and in subsequent data migrations and integrations. Making these ambiguities or contradictions visible may make them easier to address.

2.4 Prior work on Euler/X and its application to knowledge organization

In KOSs, taxonomies are hierarchies that group objects that have similar traits together (Hodge 2000). Euler/X (https://github.com/EulerProject/) was originally designed for "taxonomy" alignment—where all concepts in the taxonomies are connected via the hierarchical "is-a" relationships (Thau and Ludäscher 2007; Thau, Bowers, and

Ludäscher 2008). It is an open source tool that uses region connection calculus (RCC-5) as a reasoning tool to compare and reconcile different taxonomies. We note that other mathematical approaches have been used to align taxonomies and to monitor taxonomy evolution (e.g., Roth and Bourgine 2006; Jung 2006). Roth and Bourgine employ an approach based on Galois lattices to describe evolving, overlapping taxonomies. Similarly, the use cases driving the original development of Euler/X have been evolving, overlapping biological taxonomies (e.g., Franz, Chen, et al. 2016). In the latter approach, a domain expert asserts explicit RCC-5 articulation relationships (congruence, inclusion, overlaps, etc.) to model changes between taxonomies. Though we do not address this here, in future work we plan to explore whether (and if so, how) approaches based on Galois lattices and related approaches such as FCA (formal concept analysis), i.e., extensional approaches that make use of classes and properties to infer concept hierarchies, can be combined with intensional approaches such as Euler/X that explicitly assert hierarchy and other concept relations.

As briefly mentioned above, Euler/X can solve taxonomy alignment problems of the form $T_1 + T_2 + A \rightsquigarrow T_3$, i.e., where given taxonomies T_1, T_2 are linked via input articulations A, to produce a combined or merged solution T_3. The articulations A might be generated by a human expert or from another tool, e.g., for schema matching (Shvaiko and Euzenat 2005) or ontology matching (Euzenat and Shvaiko 2013). Sometimes the logical constraints resulting from $T_1 + T_2 + A$ are not satisfiable, so no solution (referred to as a "possible world," or PW) for T_3 exists. In other cases, the input constraints may be underspecified and the ambiguity inherent in the particular input $T_1 + T_2 + A$ allows multiple solutions for T_3, i.e., multiple possible worlds. In biological taxonomies, there is a propensity toward synthesis—finding a single tree, i.e., a single PW that reflects the ground truth. This is difficult or often impossible, if only a single vocabulary is to be used. In contrast, in Euler/X, different given vocabularies (i.e., input taxonomies T_1 and T_2) can often be reconciled into a single combined vocabulary T_3 that preserves and interrelates its constituent vocabularies T_1 and T_2. Occasionally, there are logical inconsistencies (no PWs) or ambiguities ($\geq$ 2 PWs) in the input articulations A, in which case Euler/X can help debug the former or explore and refine the latter. Usually, the main goal is to find a unique or a small number of PWs where there is no ambiguity or where it is possible to resolve the ambiguities. In either case, by finding all pairwise relationships between different taxonomies or schemas (modeled as taxonomies), Euler/X supports the reconciliation of different taxonomic perspectives.

Euler/X has been successfully applied to the problem of aligning and reconciling multiple biological taxonomies (Franz, Pier, et al. 2016; Franz, Chen, et al. 2016). More recently, the use of Euler/X for other, non-biological taxonomies has been explored with promising results (Cheng et al. 2017). The application of Euler/X to KOSs and schema may be relevant to database migration because of the many ways in which database schemas resemble or can be modeled as hierarchical structures.

3.0 Dataset: subsets of Specify schemas 1.0 & 2.3

We used Euler/X to compare two versions of the Specify database schema. As noted above, Specify is a popular biological collections management database. It was originally developed in the 1980s by the University of Kansas Biodiversity Institute (KUBI), and has been maintained by KUBI through a series of National Science Foundation grants, with the goal of transitioning to a non-profit community-driven funding model in the near future ("Specify in Transition" 2017). Over 500 museum collections use Specify software (http://www.sustain.specifysoftware.org/about/); these collections are from a range of disciplines, though the majority are biological collections (e.g., collections of animal specimens, as opposed to geological or paleontological specimens).

Specify is one of several "off-the-shelf" relational database systems designed for use with NHM collections. Each of these systems have unique database schema (e.g., Arctos has a different database structure than Specify). Though schema changes may occur in many of these systems, we chose to study Specify's schema changes for this paper, because they have been consistently documented on their website since at least 2008, and have consequently been archived by the Internet Archive (see "Specify 6 Schema" 2009). Consequently, it is an excellent case study of NHM database schema migration.

Specify's original schema (version 1.0) included one hundred thirty-eight tables ("Specify 6 Schema" 2009); the most current version (version 2.3) includes one hundred sixty-five tables ("Specify DB Schema 2.3" 2016). In general, tables have been added to either improve the database's performance and structure or to respond to changing user needs ("Documentation" 2017). In some cases, fields have been moved from one table to another. The steps we took to map and compare these two schemas are described below.

4.0 Method: mapping schemas and generating "possible worlds" with Euler/X

To compare versions of Specify Schemas with Euler/X, we first selected a subset of the Specify schemas to compare. We then mapped known relationships between at-

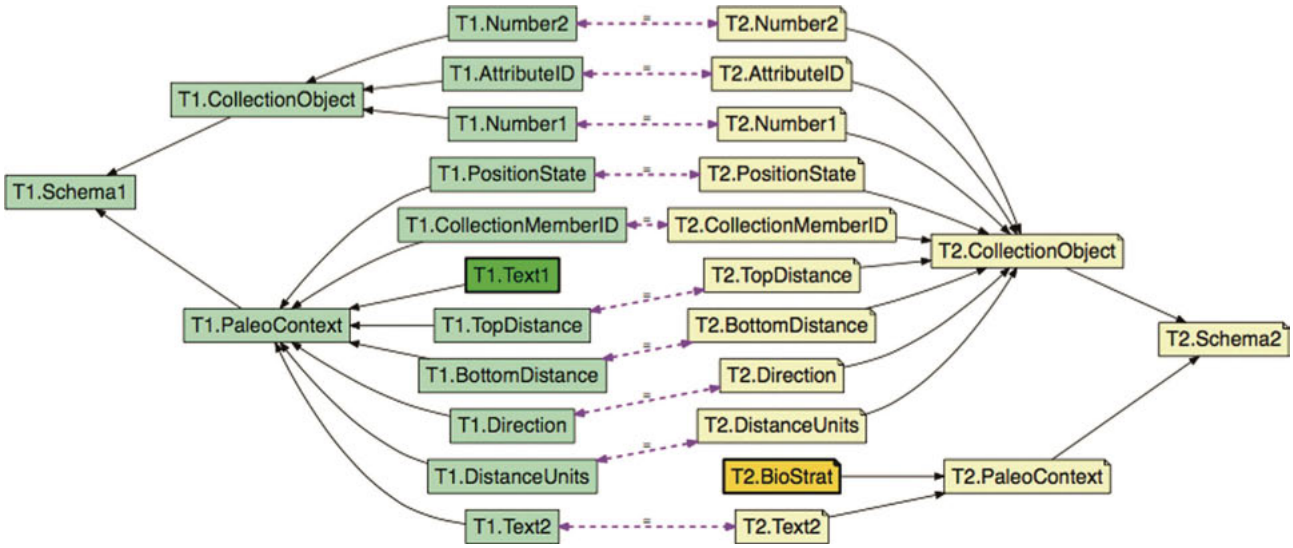

Figure 1. Visualization of input articulations between Specify Schema Version 1 (T1, left) and Specify Schema Version 2.3 (T2, right). We mapped equivalences between nodes with the same name, but left the relationship between T1.Text1 and T2.BioStrat blank, because their relationship was unclear.

tributes in the two versions of the subset Schema. We then ran our analysis. Each of these steps is described in further detail below.

4.1 Selecting a subset of the schemas

The underlying constraint problems solvable by Euler/X are computationally hard. Satisfiability of RCC-5 reasoning problems is NP-complete, which in practice can mean exponentially growing runtimes for some reasoning problems. Though Euler/X continues to evolve and improvements are being made (e.g., by reduction to different, possibly simpler underlying reasoning problems, the current prototype can run into performance issues, in particular for novice users and/or on large input problems. Consequently, we had to select a fairly small subset of the Specify schema for our initial experiments. Through research conducted in another ongoing project studying database evolution within NHMs, we learned that, over time, Specify developers have had to change the way in which contextual geological data is stored. Specifically, they have changed their approach to documenting stratigraphy in response to feedback from the paleontological community (Specify Software Project Staff 2009). In comparing Specify Schema Versions 1.0 and 2.3 we found that several attributes had been moved from the "PaleoContext" table to "CollectionObject" table. We consequently selected these two tables for comparison in Euler/X. We further selected a core subset of fields within each table for comparison (see https://github.com/akthom/EulerX-MuseumKO for the full contents of each table as well as the subsets we used for this study). We refer to Specify Schema Version 1.0 as T1 (for Taxonomy 1) and Version 2.3 as T2 (for Taxonomy 2).

4.2 Mapping attributes using the RCC-5 relations

After selecting a subset of tables and fields to compare, we reviewed the attributes of each table and aligned attributes that shared the same name. For example, in the T1.PaleoContext table, there is an attribute named "Bottom Distance." This attribute also appears on the T2.CollectionObject Table; we have mapped them as equivalent given that they share the same, unique attribute name, and given our knowledge of how the Specify schema evolved over time (acquired both through Specify's documentation and through on-going collaborations with the NHM community).

The PaleoContext table in schemas T1 and T2 each include one attribute that did not appear in the other; T1 includes a field called "Text1," and T2 includes a field called "BioStrat." While it is possible that the Specify developers simply renamed T1.Text1 as "BioStrat," we did not assume these fields to be equivalent. Instead, we rely on Euler/X to show us the ways that these fields "might" be mapped to one another through the generation of "possible worlds."

These mappings (also referred to as articulations) were input into a text file along with T1 and T2, which were then used as an input for Euler/X (See Appendix 1 or https://github.com/akthom/EulerX-MuseumKO). Nodes that are in green are from T1, and nodes that are yellow are from T2. Both T1 and T2 have fourteen nodes. The black arrows denote an "is-a or part-of" relationship between child and parent nodes within each taxonomy;

the purple dotted lines between T1 and T2 are the articulations, with "equal" signs, showing the equality relationship between the concepts. In our case, we have ten purple dotted lines, which means that our input file has ten articulations that we assert to hold.

5.0 Results: using Euler/X to generate "possible worlds"

Euler/X generated a total of five "possible worlds" from our input—that is, five alternative ways that T1 (Version 1) and T2 (Version 2.3) of the Specify Schema can be reconciled into a single "taxonomy," i.e., a combined knowledge organization comprising both schemas. We present and discuss each of these "possible worlds" below (Figures 2 to 6), and discuss the dynamics between the attributes T1 and T2 at the attribute, table, and the schema levels.

Grey boxes show where the two concepts in the two taxonomies are "congruent"—Euler/X deduced that they are exactly the same. Black arrows again showing "hierarchical" (i.e., "is-a" or "part-of") relationships within the merged taxonomy. Solid red lines are "Euler/X-inferred" hierarchical relationships between concepts; red dotted lines are the "Euler/X-inferred overlapping" relationships.

In this first "possible world" (Figure 2), at the attribute level T1.Text1 is mapped as "directly equivalent" to T2.BioStrat; the attributes are mapped as the same regardless of their different names.

At the table level, T1.CollectionObject is mapped as "being included in" T2.CollectionObject, suggesting that T2.CollectionObject has more attributes and is therefore broader than T1.Collection Object. Conversely, T1.Paleo Context is mapped as "including" the T2.PaleoContext table in this world, meaning that T2.PaleoContext is actually narrower than T1.PaleoContext. We can also see the "Euler/X-inferred overlaps" (the red dotted lines) between T1.PaleoContext and T2.CollectionObject, meaning that some of the attributes that used to be in T1.PaleoContext have been moved to T2.CollectionObject.

At the schema level, this "possible world" marked Versions 1 and 2 as equivalent; though names have changed, the fundamental structure of the schema has not.

In the second "possible world" (Figure 3), at the attribute level, T1.Text1 is mapped as "disjoint" from T2.BioStrat; that is, they are two distinct entities that neither include one another nor overlap. At the table level, T1.CollectionObject is still narrower than T2.Collection Object; however, the PaleoContext tables in T1 and T2 "overlap" with each other, meaning that they share some of the attributes, and it is unclear which is broader or narrower. At the schema level, the two versions also have an "overlapping" relationship. This overlap results from the inferred relationship of T2.CollectionObject table being totally "included in" Version 1 of the schema. Versions 1 and 2.3 of the Specify Schema, then, are overlapping but different.

In the third "possible world" (Figure 4), T1.Text1 is mapped as being "included in" T2.BioStrat. Therefore, T1.Text1 represents a subset of T2.BioStrat. At the table level, this "possible world" is similar to that in "possible world" 2 (Figure 3). However, at the schema level, Euler/X infers that Schema 1 is a subset of Version 2. In other words, Version 2.3 "includes" everything in Version 1, and thereby is an expansion of Version 1.

In the fourth "possible world" (Figure 5), at the attribute level T1.Text1 is mapped as "including" T2.BioStrat. T2.BioStrat therefore represents a subset of T1.Text1. At the table level, this "possible world" is similar to "possible world" 1 (Figure 2), in that T1.CollectionObject is included in T2.CollectionObject, and T1.PaleoContext includes T2.PaleoContext. However, at the schema level it is quite different from "possible world" 1, and the opposite of "possible world" 3 (Figure 4). In "possible world" 4, everything in Version 2.3 "is included in" Version 1; Version 2.3 thereby represents an edited or refined schema compared to Version 1.

Finally, in the fifth "possible world" (Figure 6), at the attribute level, T1.Text 1 is mapped as "overlapping" with T2.BioStrat. The two attributes share some members but not in a subset or superset relation. At the table level, it is also similar to our previous "possible worlds," in that T1.CollectionObject "is included in" T2.CollectionObject, however, the relationship between the PaleoContext tables is overlapping. At the schema level in "possible world" 5, the two versions of the schema overlap as in "possible world" 2 (Figure 3) but to a greater degree.

6.0 Discussion

6.1 Euler/X as a tool for KOS migration

Euler/X allows us to infer and then visualize all the possible relationships between two ambiguously related attributes in two versions of a database schema. The five "possible worlds" generated by Euler/X additionally show how this ambiguity propagates upward to the schema overall; the ways in which the attributes are mapped together change the ways in which the schemas overall can be mapped together. Although some of the relations between the concepts in each schema are still underspecified, Euler/X presents the five possible ways in which they could be reconciled and, thereby, could be migrated.

In the opening of this paper, we described the two issues in NHM database migration that we aimed to ad-

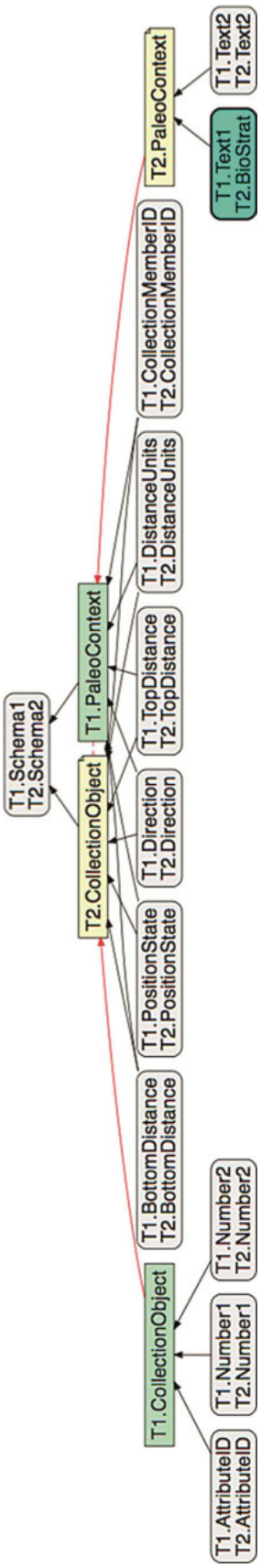

Figure 2. "Possible world" 1.

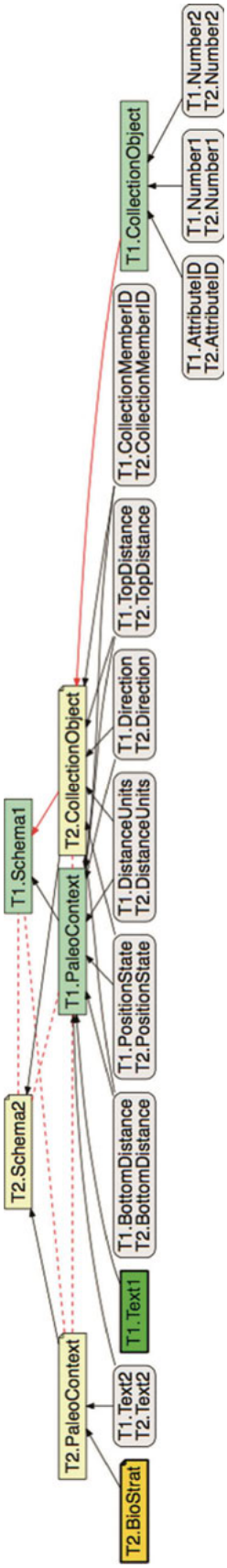

Figure 3. "Possible world" 2.

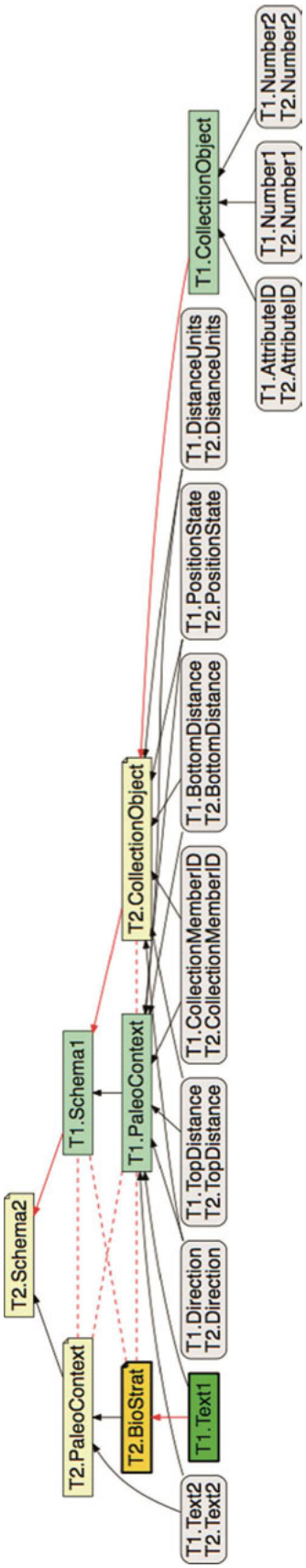

Figure 4. "Possible world" 3.

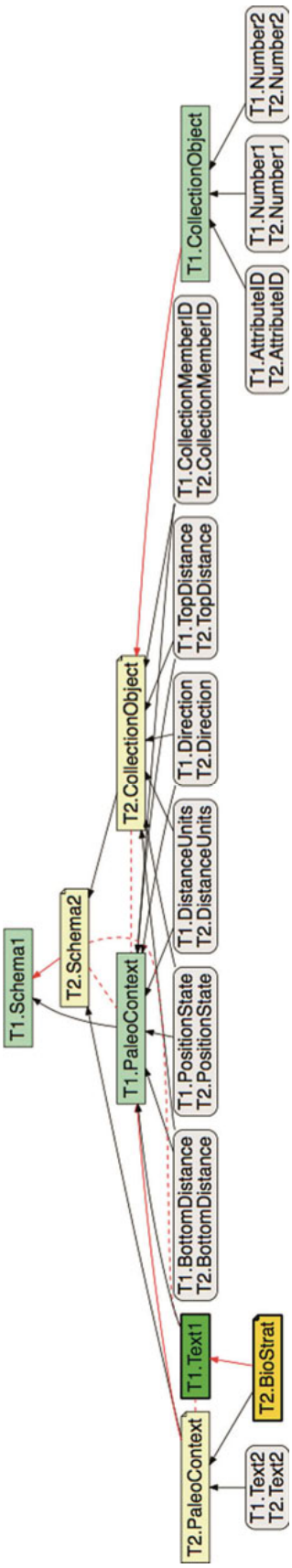

Figure 5. "Possible world" 4.

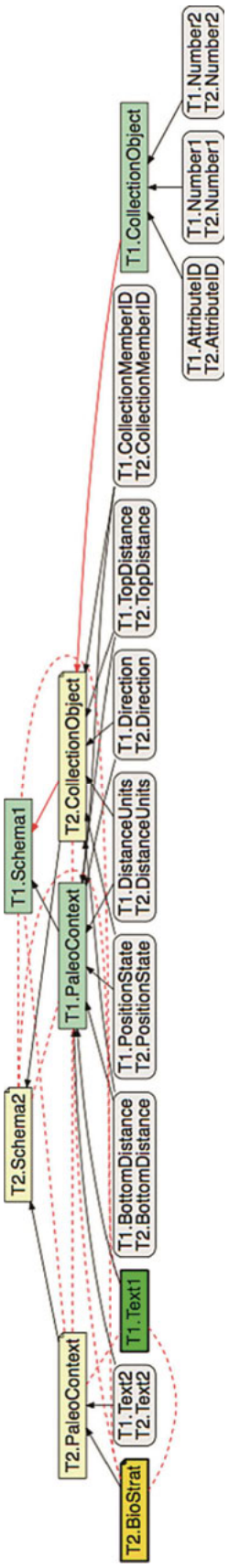

Figure 6. "Possible world" 5.

dress in this work: the need to compare old and new versions of the "same" schema; and, the need to show how aberrant use of a data model, such as the co-opting of a field for other-than-its-intended purpose, might impact a database migration process. The example presented above is an example of the first issue: the need to compare old and new versions of the same schema. Specify Version 1 (T1) includes a field called "Text1" whereas Specify Version 2.3 (T2) does not, and instead includes a field called "Biostrat." The five "possible worlds" generated by Euler/X show the five possible ways that these two fields could be related—however, which one of these five worlds is "correct" would need to be further determined by the Specify developers or the Specify user. Does their instance of Specify use T1.Text1 to store information about BioStratigraphy? If so, the first "possible world" (Figure 2) in which the two attributes are mapped as equivalent would be correct. Does their instance of Specify use T1.Text1 to store information about Biostratigraphy for some records but not in others? Then the fifth "possible world," in which the attributes are mapped as overlapping would be correct. Though Euler/X's current incarnation leaves these interpretations to the user, we can imagine Euler/X being incorporated into a database management system as a sort of "wizard" through which the user could be coached through these questions during a migration process.

Addressing the issue of data schema versioning allows us to obliquely address the issue of aberrant data model use. When a user co-opts a field, they effectively alter the semantics of the data model and thereby create a new instance or version of the data model. The same method employed above to compare two "official" versions of a data model could be employed to compare an instance of a data model as designed by a software developer, versus as deployed by an end-user. Thus, Euler/X can be used to not only show the relationship between the two different data schemas, but also changes in the "use" of two schemas. In this example, we mapped two attributes with different names as being ambiguously related, and all attributes with the same name as equivalent. However, if we were aware that, say, a data manager had used T1.Text2 to store two kinds of data, we could rerun this analysis modeling T1.Text2 and T2.Text2 as being ambiguously related as well. Thus, Euler/X can be used to help make the ramifications of aberrant or idiosyncratic use of data standards more explicit by showing all the possible logical relationships between a schema-as-originally-designed and a schema-as-it-is-used.

To this latter point, we believe that this approach may be particularly useful for planning and/or guiding the migration of a KOS such as Specify, which is built on a predetermined schema yet must sometimes be used in idiosyncratic ways by their users. As briefly reviewed above, Specify users have at times had to co-opt database attributes for local needs; as relationships between tables were changed, or database attributes renamed or moved from one table to another, the databases effectively broke and lost some of their functionality until the mappings could be repaired. We believe that Euler/X's logic-based approach could be a useful way of visualizing and disentangling these ambiguous or aberrant mappings. Euler/X could potentially even be prospectively used to show the ambiguities that may arise from changes to or changes in the use of a schema prior to the implementation of those changes.

6.2 Supporting a plurality of KOS schemas

Despite both developers' and users' best intentions, databases are often not used as their developers intend. Additionally, the breadth of legacy data structures and practices in natural history means that NHM collections data will likely always necessitate a range of different data structures, and therefore different database systems. However, the need to share data globally, as well as the need to take the burden of database design off of data curators and collections managers means that there will still be a need for centralized systems and standardized data models. Thus, individual users will likely either have to continue adapting databases to local and legacy data structures and needs through aberrant use of data attributes—or new kinds of KOSs that support a plurality of KOS schemas—potentially even within a single KOS—will need to be developed.

We believe that the approach taken here may represent a step toward supporting a plurality of KOS schemas and support of usage of a schema in multiple ways. In generating "possible worlds," Euler/X does not dictate which one should be used; rather it makes the ramifications of different data model uses and mappings visible. Examination of these "worlds" prior to database migration may prevent aberrant schema use from "breaking" a system. Further, the process of mapping two schemas together for analysis in Euler/X may help make normally tacit data practices more explicit.

We expect that co-opting database fields or otherwise using a data model in an aberrant way is a common and necessary compromise between using a well-maintained, standardized KOS and catering to idiosyncratic local needs; we further expect that this behavior is neither limited to Specify nor NHMs. In the past, we have observed that database fields are often used in ways that might make their designers cringe, particularly over time: attributes are lumped together or split apart in response to

changing needs; exceptions are made to cataloging rules and controlled vocabularies for special cases; in-house data practices need to be accounted for in unpredictable ways; and data practices evolve over time—often faster and more unpredictably than a software platform can account for or respond to. We have further found that such appropriations may inform future schema evolution (Twidale and Jones 2005) or lead to the database "learning" from its users and thereby changing shape in unexpected ways (Thomer and Twidale 2014). We argue that there is a clear need to plan (and design) for this behavior from the start, rather than only at the point of migration. We imagine that tools rooted in the same logic-based reasoning as Euler/X could be integrated into KOSs and allow users to create extremely thorough maps of their particular uses of a database over time. The logic-based approach is particularly powerful, because it could potentially be used to automate certain kinds of migrations.

7.0 Conclusion and future work

Here we have shown how Euler/X, a logic-based taxonomy alignment tool, can be used to visualize the different ways database schemas can be brought into alignment. We demonstrated this approach using a subset of two versions of the Specify database schema. We found that this approach may be helpful in KOS migration, particularly when the relationship between the old schema and the new is ambiguous, or in cases where attributes in the old schema have been co-opted or otherwise used in other-than-standard ways to meet local needs. The Euler/X approach can help make the consequences of these changes clear prior to a migration.

In our future work, we plan to continue exploring how Euler/X can be used to compare different kinds of taxonomies. Euler/X was originally designed for the comparison of biological taxonomies, which can be described as a kind of containment hierarchy—that is, "is-a" relations. Database schemas, however, are often better modeled as "part-of" hierarchies (see Varzi 2006; Keet and Artale 2008). In the study presented in this paper, we have blurred this conceptually important distinction. Indeed, the underlying RCC calculus relations can be interpreted as either is-a or part-of relationships and yield consistent results in both cases. Nevertheless, it is also clear that careful modeling of these hierarchical relationships is required to obtain meaningful inference results. In future work, we will study additional examples and alternative modeling approaches to identify new opportunities but also challenges and limitations in reasoning about schemas using RCC-based approaches.

Within the NHM KOS domain, we will expand this study to look at crosswalks between further subsets of Specify schemas or potentially to look at crosswalks between two different NHM databases such as Arctos and Specify. We believe that Euler/X could be a useful tool for making tacit data practices more explicit prior to a migration. We plan to further explore how Euler/X can be used to make tacit, in-house data practices more explicit prior to a migration, or even use Euler/X to prospectively model how non-standard uses of a database might effect migrations down the road.

References

ASC (Association of Systematics Collections, Committee on Computerization and Networking). 1993. "An Information Model for Biological Collections: Report of the Biological Collections Data Standards Workshop." http://cool.conservation-us.org/lex/datamodl.html

Berendsohn, Walter G., Anastasios Anagnostopoulos, Gregor Hagedorn, Jasmin Jakupovic, Pier Luigi Nimis, Benito Valdés, Anton Güntsch, Richard J. Pankhurst, and Richard J. White. 1999. "A Comprehensive Reference Model for Biological Collections and Surveys." *Taxon* 48:511-62. doi:10.2307/1224564

Berents, Penny, Michelle Hamer, and Vishwas Chavan. 2010. "Towards Demand Driven Publishing: Approaches to the Prioritisation of Digitisation of Natural History Collections Data." *Biodiversity Informatics* 7, no. 2. doi:10.17161/bi.v7i2.3990

Bowker, Geoffrey C. 2000. "Biodiversity Datadiversity." *Social Studies of Science* 30: 643–83. doi:10.1177/030631200030005001

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things out: Classification and Its Consequences*. Inside Technology. Cambridge, MA: MIT Press.

Brahmia, Zouhaier, Fabio Grandi, Barbara Oliboni and Rafik Bouaziz. 2015a. "Schema Versioning." *Encyclopedia of Information Science and Technology*, ed. Mehdi Khosrow-Pour, 7651-61. 3rd. ed. IGI Global. https://www.igi-global.com/chapter/schema-versioning/112468.

Brahmia, Zouhaier, Fabio Grandi, Barbara Oliboni and Rafik Bouaziz. 2015b. "Schema Evolution." In *Encyclopedia of Information Science and Technology*, 3rd ed, ed. Mehdi Khosrow-Pour. Hershey, PA: Information Science Reference, 7641-50. doi:10.4018/978-1-4666-5888-2.ch753

Brenskelle, Laura Marie. 2015. "The Use of Modern Digital Technology to Store and Serve Biodiversity Data for Research and Educational Purposes." Master's thesis, University of Texas. doi:10.15781/T21C2X

Callery, BG. 1999. "Common Names: Cooperative Access to Databased Natural History Information." In *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, ed. Mary Ellen Bowden, Trudi Bellardo Hahn and Robert V. Williams.

ASIS monograph series. Medford, NJ: Information Today, Inc. for the American Society for Information Science and the Chemical Heritage Foundation, 84-93. http://www.chemheritage.org/explore/ASIS_documents/ASIS98_main.htm

Cheng, Yi-Yun, Nico Franz, Jodi Schneider, Shizhuo Yu, Thomas Rodenhausen, and Bertram Ludäscher. 2017. "Agreeing to Disagree: Reconciling Conflicting Taxonomic Views Using a Logic-Based Approach." In *Proceedings of the 8th Annual SIS&T Meeting, Washington, D.C., 27 October-1 November, 2017*. https://www.ideals.illinois.edu/handle/2142/97907

"Documentation." 2017. At *Specify Software Project* website. http://specifyx.specifysoftware.org/documentation/

Dubin, David, Joe Futrelle, Joel Plutchak, and Janet Eke. 2009. "Preserving Meaning, Not Just Objects: Semantics and Digital Preservation." *Library Trends* 57, no. 3:595-610.

Euzenat, Jérôme, and Pavel Shvaiko. 2013. *Ontology Matching*. Berlin: Springer. doi:10.1007/978-3-642-38721-0

Franz, Nico M., Mingmin Chen, Parisa Kianmajd, Shizhuo Yu, Shawn Bowers, Alan S. Weakley, and Bertram Ludäscher. 2016. "Names Are Not Good Enough: Reasoning over Taxonomic Change in the Andropogon Complex[1]." *Semantic Web* 7: 645-67. doi:10.3233/SW-160220

Franz, Nico M., Naomi M. Pier, Deeann M. Reeder, Mingmin Chen, Shizhuo Yu, Parisa Kianmajd, Shawn Bowers, and Bertram Ludäscher. 2016. "Two Influential Primate Classifications Logically Aligned." *Systematic Biology* 65:561-82. doi:10.1093/sysbio/syw023

Galante, Renata de Matos, Clesio Saraiva dos Santos, Nina Edelweiss, and Álvaro Freitas Moreira. 2005. "Temporal and Versioning Model for Schema Evolution in Object-Oriented Databases." *Data & Knowledge Engineering* 53:99-128. doi:10.1016/j.datak.2004.07.001

Gao, Shi, and Carlo Zaniolo. 2012a. "Provenance Management in Databases Under Schema Evolution." Paper presented at TaPP'12: 4th USENIX Workshop on the Theory and Practice of Provenance. https://www.usenix.org/system/files/conference/tapp12/tapp12-final20.pdf

Gao, Shi, and Carlo Zaniolo. 2012b. "Supporting Database Provenance Under Schema Evolution." In Advances *in conceptual modeling: ER 2012 workshops: CMS, ECDM-NoCoDA, MoDIC, MORE-BI, RIGiM, SeCoGIS, WISM, Florence, Italy, October 15-18, 2012, proceedings* ed. Silvana Castano, Panos Vassiliadis, Laks V. S. Lakshmanan, and Mong Li Lee,. Lecture Notes in Computer Science 7518. Berlin: Springer-Verlag, 67-77. doi:10.1007/978-3-642-33999-8_9

Gnoli, Claudio. 2008. "Ten Long-Term Research Questions in Knowledge Organization." *Knowledge Organization* 35:137-49.

Grinnell, Hilda W. 1958. *Annie Montague Alexander*. Berkeley, CA: Grinnell Naturalists Society.

Harping, Patricia. 2014. "Introduction to Metadata: Crosswalk," revised 9 June 2009 by Patricia Harpring; ed. Murtha Baca, Patricia Harpring, Jon Ward, and Antonio Beecroft. Compiled by Murtha Baca, Sherman Clarke, Jan Eklund, Anne J. Gilliland, Patricia Harpring, Mary S. Woodley, and Elizabeth O'Keefe. J. Paul Getty Trust. http://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html

JISC. 2009. "Vocabulary Mapping Framework (VMF): An Introduction v1.0." http://www.doi.org/VMF/documents/VocabularyMappingFrameworkIntroductionV1.0(091212).pdf

Jung, Jason J. 2006. "Taxonomy Alignment for Interoperability Between Heterogeneous Digital Libraries." In *Digital Libraries: Achievements, Challenges and Opportunities*, ed. Shigeo Sugimoto, Jane Hunter, Andreas, Rauber Atsuyuki, and Morishima, 274–82. Lecture Notes in Computer Science 4312. [New York, N.Y.?]: Springer doi:10.1007/11931584_30

Keet, C. Maria, and Alessandro Artale. 2008. "Representing and Reasoning over a Taxonomy of Part–whole Relations." *Applied Ontology* 3: 91–110.

Khoo, Michael, and Catherine Hall. 2010. "Merging Metadata: A Sociotechnical Study of Crosswalking and Interoperability." In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. New York: ACM, 361-4. doi:10.1145/1816123.1816180

Lauruhn, Michael and Paul Groth. 2016. "Sources of Change for Modern Knowledge Organiza$on Systems." *Knowledge Organization* 43:622-29.

"Metadata Schema Transformation Services." 2014. At *OCLC Research* website. http://www.oclc.org/research/themes/data-science/schematrans.html

Perrine, John D., and James L. Patton. 2011. "Letters to the Future." In *Field Notes on Science & Nature*, ed. Michael R. Canfield. Cambridge, Mass: Harvard University Press, 211-50.

Roddick, John F. 1992. "Schema Evolution in Database Systems: An Annotated Bibliography." *SIGMOD Record* 21, no. 4:35-40. doi:10.1145/141818.141826

Roth, Camille, and Paul Bourgine. 2006. "Lattice-Based Dynamic and Overlapping Taxonomies: The Case of Epistemic Communities." *Scientometrics* 69:429-47. doi:10.1007/s11192-006-0161-6

Scharnhorst, Andrea, Richard P. Smiraglia, Christophe Guéret and Alkim Almila Akdag Salah. 2016. "Knowledge Maps of the UDC: Uses and Use Cases." *Knowledge Organization* 43:641-54.

Shvaiko, Pavel, and Jéróme Euzenat. 2005. "A Survey of Schema-Based Matching Approaches." In *Journal on Data Semantics IV*, ed. Stefano Spaccapietra. Lecture Notes in Computer Science 3730. Berlin: Springer, 146-71. doi:10.1007/11603412_5

"Specify 6 Schema." 2009. https://web.archive.org/web/20090204133612/http://specify6.specifysoftware.org:80/SpecifySchema.html

"Specify DB Schema 2.3." 2016. Updated 3 November 2016. http://www.specify6.specifysoftware.org/schema.html

"Specify in Transition." 2017. At *Specify* website. http://www.sustain.specifysoftware.org/transition/

Specify Software Project Staff. 2009. "Specify 6's Approach to Stratigraphy: The Specify 6 Paleontological Data Model, 31 March 2009, version 2.0." https://web.archive.org/web/20131207174733/http://specifysoftware.org/content/specify-6s-approach-stratigraphy.

St. Pierre, Margaret, and William P. LaPlant. 1998. "Issues in Crosswalking Content Metadata Standards." Washington, D.C.: NISO. http://www.niso.org/publications/white_papers/crosswalk/

Tennis, Joseph T. 2012. "The Strange Case of Eugenics: A Subject's Ontogeny in a Long-lived Classification Scheme and the Question of Collocative Integrity." *Journal of the American Society for Information Science and Technology* 63:1350-59. doi:10.1002/asi.22686

Tennis, Joseph T. 2016. "Methodological Challenges in Scheme Versioning and Subject Ontogeny Research." *Knowledge Organization* 43:573-80.

Thau, David, Shawn Bowers, and Bertram Ludäscher. 2008. "Merging Taxonomies under RCC-5 Algebraic Articulations." In *Proceedings of the 2nd International Workshop on Ontologies and Information Systems for the Semantic Web*. New York: ACM, 47–54. doi:10.1145/1458484.1458492

Thau, David, and Bertram Ludäscher. 2007. "Reasoning about Taxonomies in First-Order Logic." *Ecological Informatics* 2:195-209. doi:10.1016/j.ecoinf.2007.07.005

Thomer, Andrea K., Karen S. Baker, Simone Sacchi, and David Dubin. 2012. "Completeness, Coverage & Equivalence in Scientific Data Records." *Proceedings of the American Society for Information Science and Technology*, 49:1-4. doi:10.1002/meet.14504901331

Thomer, Andrea K., and Michael B Twidale. 2014. "How Databases Learn." *iConference 2014 Proceedings*, ed. Maxi Kindling and Elke Greifeneder, 827-33. doi:10.9776/14409

Trigg, Randall H., Jeanette Blomberg, and Lucy Suchman. 2002. "Moving Document Collections Online: The Evolution of a Shared Repository." In *ECSCW '99: Proceedings of the Sixth European Conference on Computer Supported Cooperative Work 12–16 September 1999, Copenhagen, Denmark*, ed. Suanne Bødker, Morten Kyng, and Kjeld Schmidt. Dordrecht: Kluwer Academic Publishers, 331-50. doi:10.1007/0-306-47316-X_18

Twidale, Michael B., and M. Cameron Jones. 2005. "'Let Them Use Emacs': The Interaction of Simplicity and Appropriation." *International Reports on Socio-Informatics* 2, no. 2:78-84.

Varzi, Achille C. 2006. "A Note on the Transitivity of Parthood." *Applied Ontology* 1:141-46.

Wieczorek, John, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. 2012. "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard," ed. Indra Neil Sarkar. *PLoS ONE* 7 (1): e29715. doi:10.1371/journal.pone.0029715

Zeng, Marcia Lei, and Lois Mai Chan. 2006. "Metadata Interoperability and Standardization: A Study of Methodology Part II; Achieving Interoperability at the Record and Repository Levels." *D-Lib Magazine* 12, no. 6. http://www.dlib.org/dlib/june06/zeng/06zeng.html

Appendix 1 Euler/X Input File

```
taxonomy T1 SpecifyT1
(Schema1 PaleoContext CollectionObject)
(PaleoContext BottomDistance CollectionMemberID Direction DistanceUnits PositionState Text1 Text2 TopDistance)
(CollectionObject AttributeID Number1 Number2)

taxonomy T2 SpecifyT2
(Schema2 PaleoContext CollectionObject)
(PaleoContext BioStrat Text2)
(CollectionObject AttributeID BottomDistance CollectionMemberID Direction DistanceUnits PositionState TopDistance Number1 Number2)

articulations T1 T2
[T1.BottomDistance equals T2.BottomDistance]
[T1.Text2 equals T2.Text2]
[T1.CollectionMemberID equals T2.CollectionMemberID]
[T1.Direction equals T2.Direction]
[T1.DistanceUnits equals T2.DistanceUnits]
[T1.Number1 equals T2.Number1]
[T1.Number2 equals T2.Number2]
[T1.AttributeID equals T2.AttributeID]
[T1.PositionState equals T2.PositionState]
[T1.TopDistance equals T2.TopDistance]
```