Practical and meta- challenges in modeling database migrations: a case study of the MBGNA

Andrea K. Thomer*, Samuel Sciolla, Kathryn Topham, Alexandria Rayburn & Maryse Lundering-Timpano
School of Information, University of Michigan, 105 S. State St.Ann Arbor, MI 48109-1285
*Corresponding author: athomer@umich.edu

Introduction: migrating research data collections

A fundamental aspects of digital collections management is *database migration:* "the process of moving data from one information system or storage medium to another to ensure continued access to the information as the system or medium becomes obsolete or degrades over time" ("migration," n.d.). But while analog/physical "data" migration is well understood and theorized within LIS, digital collection migration is less well supported — particularly with regards to the conceptual modeling work that must be done to facilitate migrations and reverse engineer legacy systems.

In the "Migrating Research Data Collections" project (IMLS Grant # RE-07-18-0118-18), we are developing these best practices through a multi-case study of database and digital collections migration. We are starting by developing case studies of database migration in natural history museums (NHMs), which are often overlooked in LIS research despite being early adopters of database technology and early contributors to scholarship in data curation (Palmer, Weber, Renear, & Muñoz, 2013; Thomer, Weber, & Twidale, 2018).

Here we present a case study being developed through this project: database migration and maintenance at the Matthaei Botanical Gardens and Nichols Arboretum (MBGNA). Over the last two years, staff at the MBGNA have been working to migrate disparate legacy data sources together into one centralized ArcGIS database that they hope will enable more tailored data entry and analysis. This case illustrates practical issues related to conceptual modeling that information professionals face during database migrations — as well as some meta-issues faced by us as researchers studying database migration.

Method

This case study was developed through semi-structured interviews (45 to 75 minutes each) with four curatorial staff at the MBGNA; meetings with the team to verify dates and findings; and close examination and comparison of three versions of their databases. Meetings and interviews took place in 2018, and analysis of interview transcripts and databases continued through 2019.

Interview transcripts were used to develop a timeline of database migrations, visualized as a Sankey diagram (Figure 1).

The three versions of the databases examined include the following:

1) a Microsoft Access database (.mdb file extension) representing the main database circa 2008
2) a Microsoft Access database (.accdb file extension) representing an interim revision of the main database schema, used to plan a migration circa 2016
3) and an online ArcGIS database last updated in December 2018.

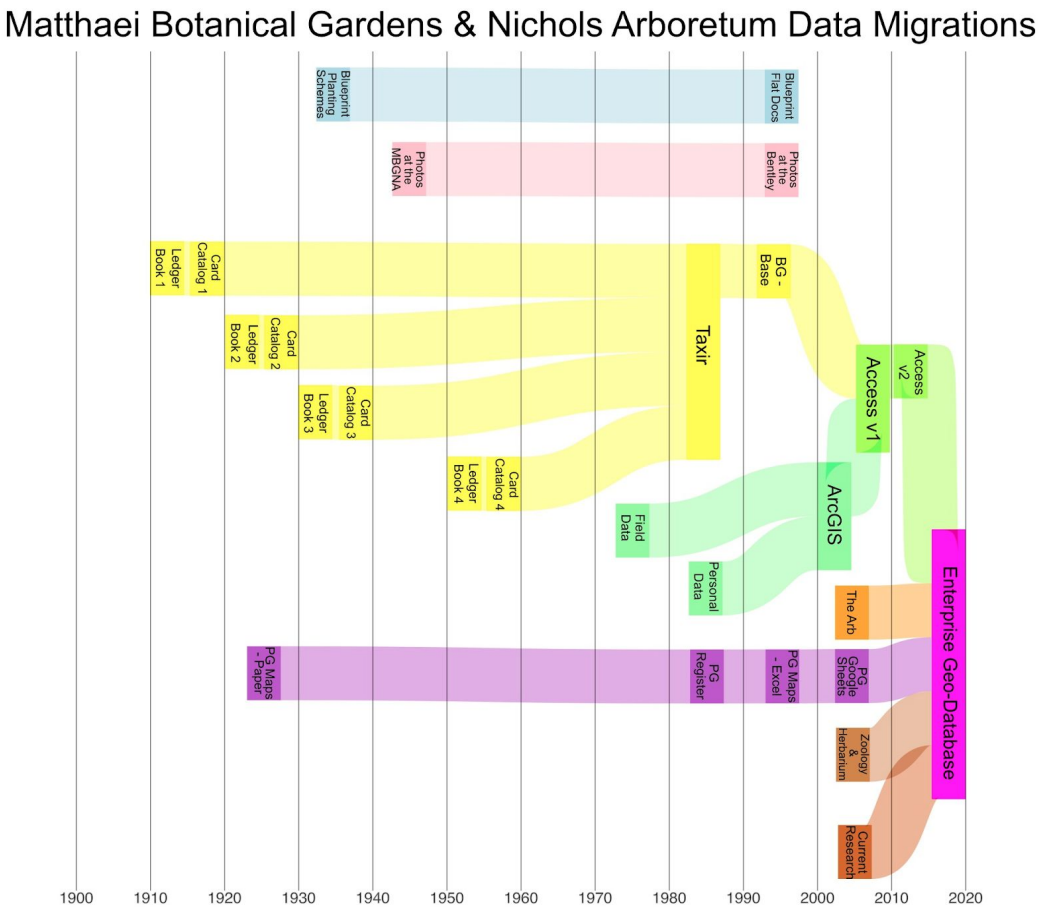
Because the three databases span multiple formats, we migrated each of them to MySQL as a sort of platform-agnostic crosswalk to facilitate comparison. Access databases were migrated with the MySQL Migration Wizard[1], and the ArcGIS database was migrated via a brief Python script[2]. Once each database's schema was migrated, we used MySQL Workbench's Reverse Engineering[3] capabilities to produce Enhanced Entity Relationship (EER) diagrams that could be arranged, corrected, and analyzed.

Case: Database migration at the MBGNA

The MBGNA's collections and catalogs date back to 1910. Early records were stored in a card catalog, in which every plant in the "living collection" (garden) was cataloged on an individual card. In the 1980s card catalogs were first transcribed into TAXIR, a data system for NHMs (Estabrook, 1979; Hudson, Dutton, Reynolds, & Walden, 1971). The database was migrated to BG-Base[4] (a platform specifically designed for botanical gardens) in the 1990s; then to Microsoft Access in 2003; and updated to the most recent version in 2012. Since our initial meetings with MGBNA staff in 2018, the Access database has been migrated to an ArcGIS "GeoDatabase" (Figure 1).

In addition to the collections catalog data, MBGNA staff have long maintained separate data stores for specific gardens (notably the Peony garden[5], which is managed as its own entity), one-off inventories of the gardens, or individual field projects. These auxiliary data stores are typically not organized according to any standard and are not always freely shared; one participant described a history of treating field data, *"as jewels of individual dragons, in terms of, 'This is my information, not yours.'"* MBGNA still maintains many of these auxiliary data collections, including the Peony Garden database — but they hope to migrate this and other data stores to ArcGIS as well.

[1] https://dev.mysql.com/doc/workbench/en/wb-migration.html
[2] https://github.com/ssciolla/mrdc_mbgna_arcgis_analysis
[3] https://dev.mysql.com/doc/workbench/en/wb-reverse-engineer-live.html
[4] http://www.bg-base.com/
[5] https://peony.mbgna.umich.edu/



(Figure 1: A Sankey diagram illustrating historical data migrations at the MBGNA. PG = Peony Garden)

The structure of the most recent Access database was based on the International Transfer Format for Botanical Gardens (ITF), a content-focused metadata standard (agnostic to encoding) for creating interoperable accession-level records that can be shared easily between institutions (Botanic Gardens Conservation International, 1998) that was developed in the 1980s for botanical gardens (Botanic Gardens Conservation Secretariat, 1987). This structure has largely been carried forward into the ArcGIS database.

Close examination of the three EER diagrams representing the most recent versions of the main catalog give a sense of a schema in transition, catalyzed in part by new standards and technologies. The oldest database (the 2008 Access database) exhibits design choices tailored to data entry and shows signs of significant use over time. For instance, the “tblplant” table features a series of Boolean columns that act as checklists indicating where the plant originated (e.g. “tblPlant_ContinentAsia”), in what months it blooms (e.g. “tblPlant_BloomJun”), and its flower color (e.g. “tblPlant_FlowerColorGreen”). Though these tables could be further normalized through many-to-many relationships, the Boolean columns (in combination with customized data entry forms) likely made data entry easier (Appendix A, Figure 1). The 2008 database also feature many tables without formal relationships to other tables; these may be artifacts of some revision process (Appendix A, Figure 2). The 2016 Access database represents an effort to refine

the schema, largely through reference to the ITF standard, and features many changes to tables driven by terms in ITF (Appendix A, Figure 3). We note that this "version" of the database was never used to actually store data, but rather acted as an interim modeling environment, wherein collections staff attempted to refactor the schema before migrating the data to ArcGIS.

Finally, the most recent database built using ArcGIS, features fewer tables than its predecessors, largely because ArcGIS does not treat controlled vocabularies (or "domains" in their parlance) as separate tables, as the prior databases did. Also, the ArcGIS platform introduces a distinction between tables and layers, in which layers are tables that contain data related to geocoordinates and shape to assist in presenting the data on a map.

Discussion: practical and "meta" issues in modeling migrations

This case exemplifies several common challenges to database migrations (and modeling database migrations), including issues in documenting changes to a database's schema, levels of normalization, and entity definitions. The changes in the database over time reflect the institution's efforts to simplify and standardize its schema: incorporating terms and entity classes from the ITF standard, and the subsequent migration and deletion of legacy tables and fields. In some cases, these changes introduce violations to normal forms — and therefore potentially introduce issues of data inconsistency. This may point to a need for resources to assist LAM staff in defining the entities and relationships in their problem space, applying normalization rules, and populating databases with normalized schemas.

The main MBGNA database's trajectory from card catalog to numerous databases is similar to that of other NHMs studied in our prior work (Thomer et al., 2018); most collection staff we previously spoke with were similarly dealing with legacy databases that had been migrated repeatedly over decades, but with little documentation showing how or why. They also had to reverse engineer poorly documented legacy data structures, and it seems that there is a lack of best practices or support to guide them in this work. Given the common task of reverse engineering preexisting data structures — and understanding changes to data structures over time — there is a clear need for structured *and accessible* ways of documenting legacy data flows.

In developing our own understanding of MBGNA's legacy data systems, we encountered challenges to modeling migrations at a somewhat meta-level: in order to analyze how staff reverse engineered and migrated databases, we had to conduct our own reverse engineering of the data models! In doing so, we identified two approaches that may be helpful to both researchers and practitioners: the EER and Sankey diagrams presented above. EER diagrams are a familiar tool to some but are inconsistently taught in iSchools and database classes. More consistent training in diagramming languages like EER and UML would be a simple way to better support database documentation and migration. The Sankey diagram above is a less conventional approach to showing relationships between different versions of databases, but may have promise in providing a big picture view of systems over time. Both of these diagramming

methods have the potential to act as the "presentation view" that Jagadish has argued is necessary to give modern database users a clear understanding of their systems (Jagadish et al., 2007).

In future work, we will be further refining these diagramming approaches, and seeking feedback from our studying participants on their efficacy in illustrating database changes over time. We also hope to draw on logic-based schema alignment to visualize changes to taxonomies over time (Chen, Yu, Franz, Bowers, & Ludäscher, 2014; Cheng et al., 2017; Franz et al., 2015; A. Thomer, Cheng, Schneider, Twidale, & Ludäscher, 2017).

Acknowledgements

This work was funded by IMLS Grant # RE-07-18-0118-18. Thanks to our study participants for their time and excellent contributions. Thanks to Daria Orlowska, Clare Michaud, and Nicco Pandolfi for early assistance in data collection and analysis of data schemas.

References

Botanic Gardens Conservation International. (1998). *The International Transfer Format (ITF) for botanic garden plant records. Version 2* (No. 2; p. 86).

Botanic Gardens Conservation Secretariat. (1987). *The International Transfer Format (ITF) for botanic garden plant records. Version 01.00* (No. 01.00). https://doi.org/10.5962/bhl.title.45427

Chen, M., Yu, S., Franz, N., Bowers, S., & Lu\''asher, B. (2014). Euler/X: A Toolkit for Logic-based Taxonomy Integration. *ArXiv:1402.1992 [Cs]*. Retrieved from http://arxiv.org/abs/1402.1992

Cheng, Y.-Y., Franz, N., Schneider, J., Yu, S., Rodenhausen, T., & Ludäscher, B. (2017). Agreeing to disagree: reconciling conflicting taxonomic views using a logic-based approach. *Proceedings of the Association for Information Science and Technology*. Presented at the The Association for Information Science and Technology, Washington, D.C. Retrieved from https://www.ideals.illinois.edu/handle/2142/97907

Estabrook, G. F. (1979). A TAXIR Data Bank of Seed Plant Types at the University of Michigan Herbarium. *Taxon*, *28*(1/3), 197–203. https://doi.org/10.2307/1219576

Franz, N. M., Chen, M., Yu, S., Kianmajd, P., Bowers, S., & Ludäscher, B. (2015). Reasoning over Taxonomic Change: Exploring Alignments for the Perelleschus Use Case. *PLOS ONE*, *10*(2), e0118247. https://doi.org/10.1371/journal.pone.0118247

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

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

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

Palmer, C., Weber, N. M., Renear, A., & Muñoz, T. (2013). *Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data*. Retrieved from https://www.ideals.illinois.edu/handle/2142/78099

Thomer, A., Cheng, Y.-Y., Schneider, J., Twidale, M., & Ludäscher, B. (2017). Logic-Based Schema Alignment for Natural History Museum Databases. *Knowledge Organization*, *44*, 545–558. https://doi.org/10.5771/0943-7444-2017-7-545

Thomer, A. K., Weber, N. M., & Twidale, M. B. (2018). Supporting the long-term curation and migration of natural history museum collections databases: Supporting the Long-term Curation and Migration of Natural History Museum Collections Databases. *Proceedings of the Association for Information Science and Technology*, *55*(1), 504–513. https://doi.org/10.1002/pra2.2018.14505501055}

Appendix A: Database EER Diagrams

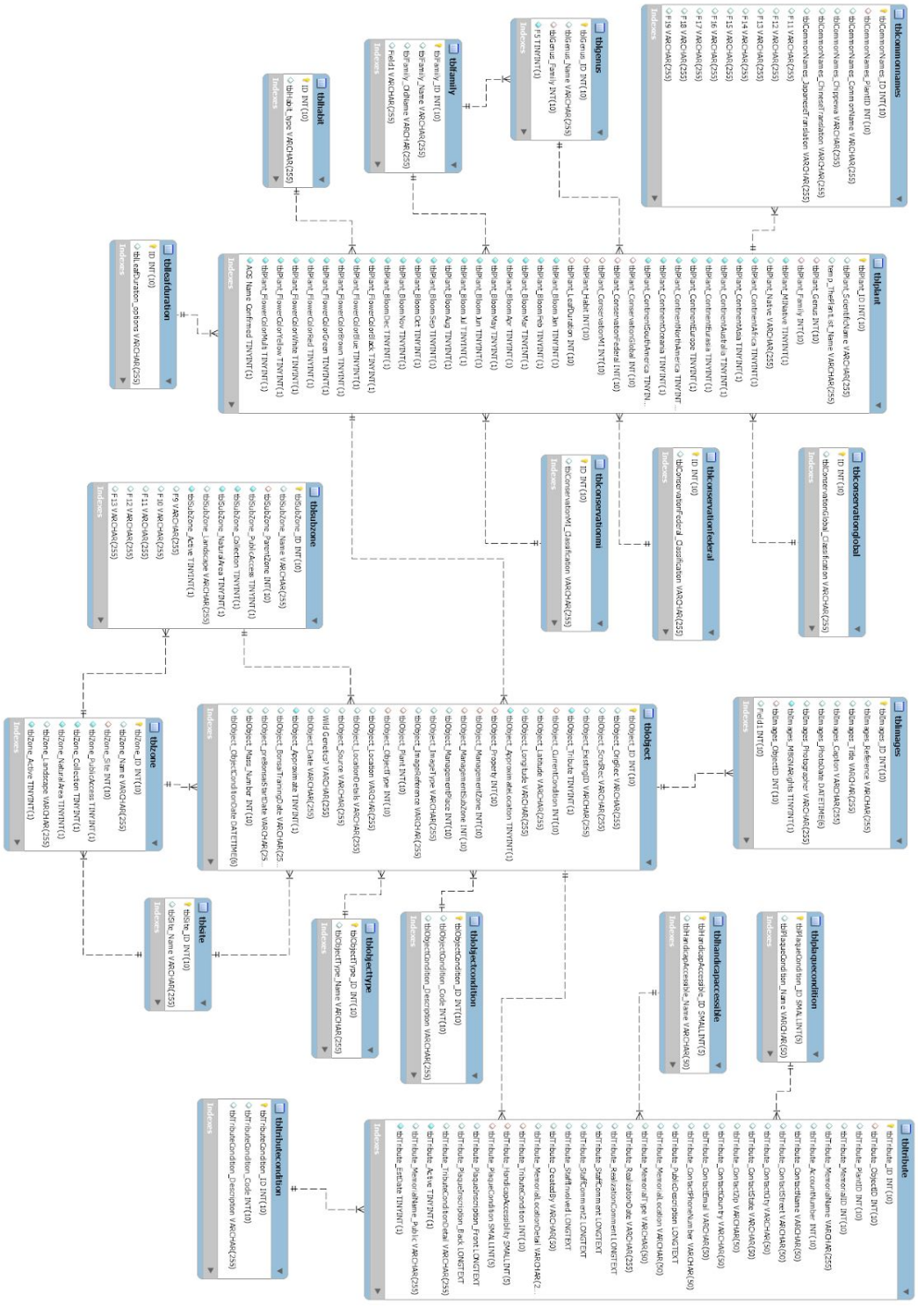


Figure 1. 2008 Database - The above EER diagram represents the schema of the first database from the MBGNA case study. To generate the above diagram, we migrated the original Microsoft Access database (.mdb file extension) to MySQL using MySQL Workbench. In addition to the Boolean fields described above, some tables feature blank columns (usually starting with “F”) that appear to be mistakenly created.

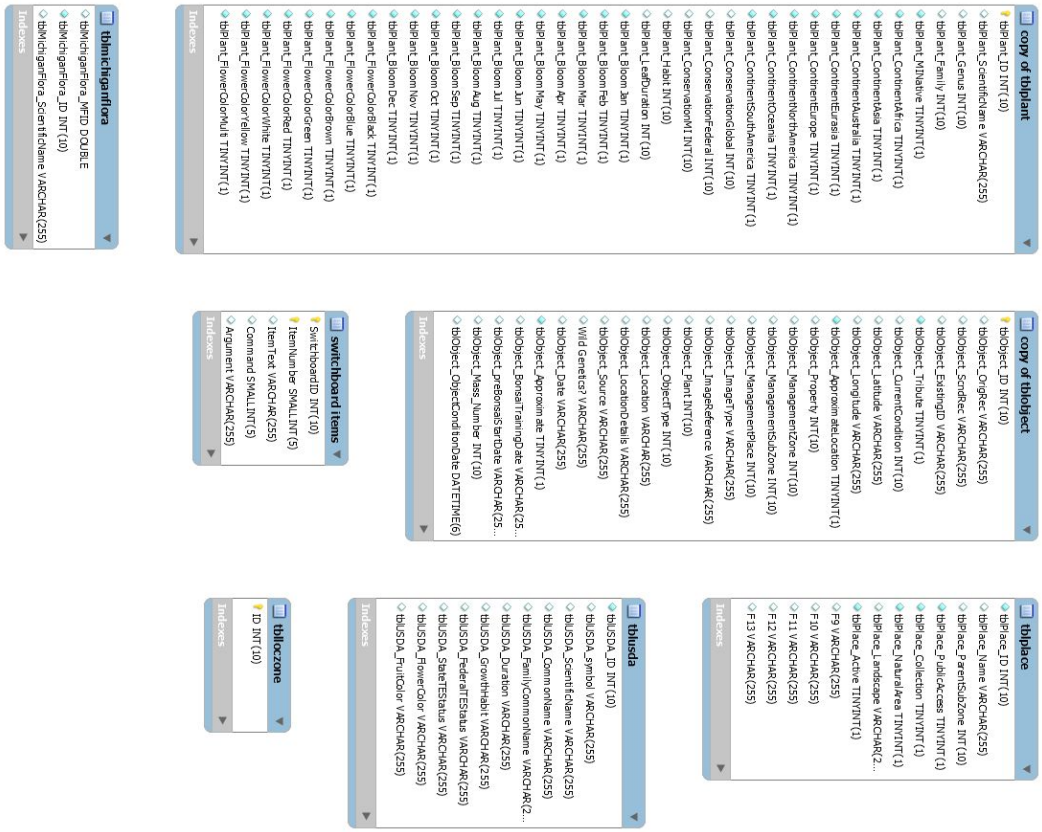


Figure 2. 2008 Database (Extra Tables) - The first Microsoft Access database from the MBGNA case study also contained the above seven tables, all of which had no declared relationships.

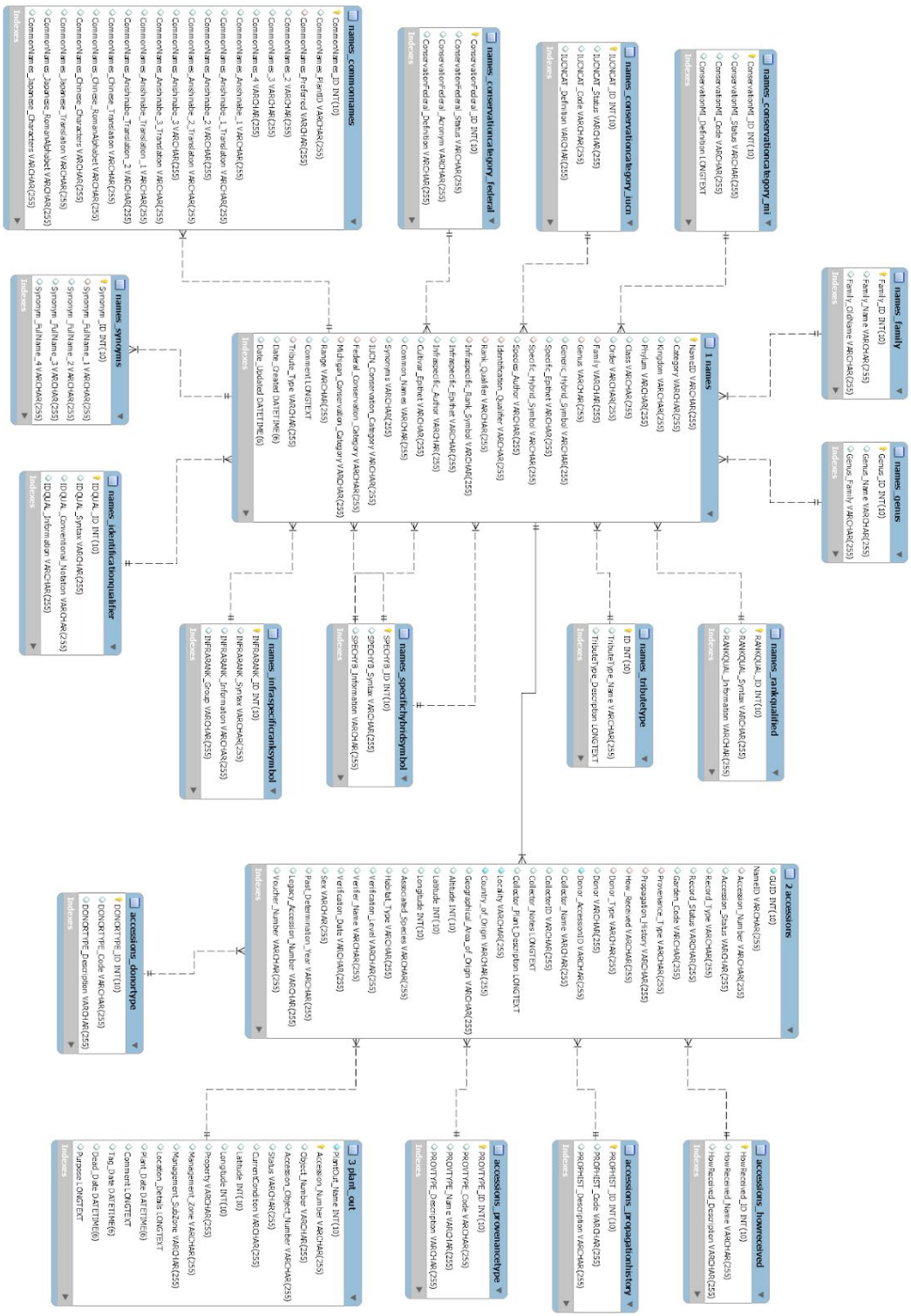


Figure 3. The 2016 Access database - The above EER diagram represents the schema of the second MBGNA database, an Access database (.accdb file extension) that was an in-progress revision of the 2008 database. As with the first, we migrated the database and created the diagram using MySQL Workbench. The database represents an effort to refine the schema, largely through reference to the ITF standard.

Though the standard specification only includes a small section on database design, it does roughly sort its fields into two categories: "accession-based" fields, serving to record a specimen or a group of specimens in the catalog, and "taxon-based" fields, serving to record data about specimen types and their associated taxonomic ranks. These two approaches manifest as the “1 names” and “2 accessions” tables (REF to figure or appendix), respectively — changed from the 2008 version’s “tblplant” and “tblobject” tables — and numerous column names come directly from the ITF fields associated with the categories (e.g. “Accession Status” in “2 accessions” table and “Cultivar Epithet” in “1 names”).

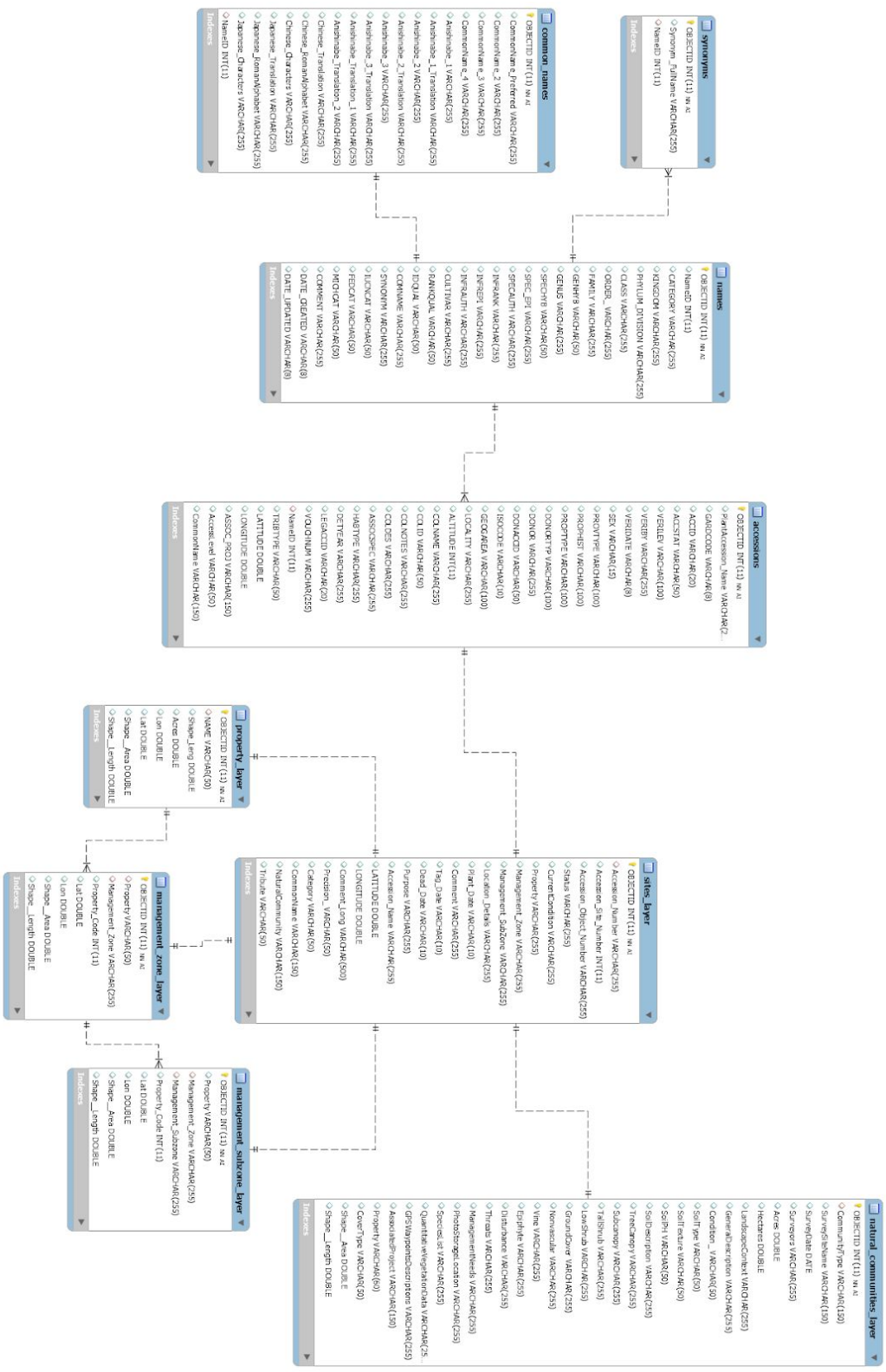


Figure 4 - ArcGIS database - The above diagram represents the schema of the third MBGNA database, which is browsable online through ArcGIS. We created the diagram by using a Python script to convert the schema details available in JSON into a SQL script, which was then processed using MySQL and

MySQL Workbench. The ArcGIS database contains fewer tables than its predecessors, largely because ArcGIS does not treat controlled vocabularies (or “domains” in their parlance) as separate tables, as the prior databases did. Also, the ArcGIS platform introduces a distinction between tables and layers, where layers are tables that contain data related to geocoordinates and shape to assist in presenting the data on a map. Interestingly, the 2008 database had an inheritance chain of zone or space tables; the 2016 database simplified them into a “3 plant out” table; and the ArcGIS database reintroduced, modified, and augmented the inheritance chain as layers. Considering concepts such as zone, subzone, and property as separate entities seems to have been useful in the ArcGIS context.