How Databases Learn

Andrea K. Thomer[1,2] and Michael B. Twidale[2]

[1] Center for Informatics Research in Science and Scholarship, University of Illinois at Urbana-Champaign

[2] Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Abstract

The relational database has been a fixture of the modern research laboratory -- used to catalog and organize specimens and petri dishes, as well as to organize and store research data and analyses. Yet, though there are numerous textbooks on database design and short-term maintenance, there is still a need for deeper exploration of how these artifacts change, grow and are maintained in the long term, and how their very structure can affect their users' work. Findings from a deeper, more extended exploration of database use over long periods of time would have implications for not just data curation, preservation and management, but also for our understanding of actual, situated information organization practices and needs in science: designing for actual practice rather than for unrealistic idealization of these practices and needs. We draw inspiration, and our title, from Brand's highly influential book; "How Buildings Learn" (1995). We believe many of the topics Brand discusses regarding buildings' change and growth over time might usefully be applied to certain aspects of databases. This work is a first step towards understanding how databases, like buildings, learn.

Keywords: databases, data migration, trace ethnography, information organization, data preservation

Citation: Thomer, A. K., & Twidale, M. B. (2014). How Databases Learn. In *iConference 2014 Proceedings* (p. 827–833). doi:10.9776/14409

Acknowledgements: Thanks to John Noyes and Matt Yoder at the Illinois Natural History Survey for providing us with access to the Universal Chalcidoidea Database database.

Contact: thomer2@illinois.edu, twidale@illinois.edu

1 Introduction

For at least the last 40 years, the relational database has been a fixture of the modern research laboratory -- used to catalog and organize specimens and petri dishes, as well as to organize and store research data and analyses; Manovich goes so far as to call them the "key form of cultural expression" in the computer age (1999). Yet, though there are numerous textbooks on database design and short-term maintenance, and a fair amount of LIS and CSCW literature exploring people's use of, and on-going collaboration around, databases, there is still a need for deeper exploration of how these artifacts change, grow and are maintained in the long term, and how their very structure can affect their users' work. Findings from this deeper, more extended exploration would have implications for not just data curation, preservation and management, but also for our understanding of actual, situated information organization practices and needs in science: designing for actual practice rather than for unrealistic idealization of these practices and needs.

1.1 How databases, like buildings, learn

We draw inspiration, and our title, from Brand's highly influential book; "How Buildings Learn" (1995). We believe many of the topics Brand discusses regarding buildings' change and growth over time might usefully be applied to certain aspects of databases. To illustrate, we list here some issues in the book that have promise as provocative analytic lenses to apply to studies of databases:

- Consideration of the database beyond its initial construction over timescales of years and decades – admitting at least the possibility that growth and use might last for centuries

- Consideration of gradual evolution of the database over time; growth, accretion of extra parts, slow changes in how it is used, accommodations to new technologies
- Periodic radical repurposing of the database, changing understanding of what it is 'for'
- Shearing: how different aspects of the design may need to change at different rates
- Consideration of the database as a resource that can be better understood by how people use it, not simply as a designed artifact that can be analyzed by looking at it without any people around
- Acknowledging that people will appropriate the database, often violating the pure design intents of its architect
- Tensions between a carefully thought out architecture and a more vernacular style of initial creation and modification
- Tensions between rigid control of form and use enforcing consistency and reliability versus more adaptable and responsive but idiosyncratic evolving use
- Contrasting a doomed attempt to design the database right in the first place with designing to make it easier to modify as circumstances change

In this paper, we present preliminary steps toward what Schuurman calls a "database ethnography" (2008), similar to Geiger and Ribes' trace ethnography (2011): a detailed examination of changes of database table structure and schema over time, via a case study detailing the migration of a relational database from one system to another. The database under study is the Universal Chalcidoidea Database -- a long-lived natural history database containing nomenclatural data about a large superfamily of wasps. This study could be used to inform both database design, database curation, and our understanding of how people collaboratively use, alter and maintain information structures to do work. Research questions guiding this work include:

- How do databases in general, and *relational* databases in particular, shape the work or research that is done with them?
- How do database structures or schemata affect the work that is done with them both at the time of their creation and long after?
- What are the recurrent dilemmas in collaborative database design, use and maintenance?

And finally,

- How does a database learn?

2 Prior Work

Prior work exploring database use over time falls primarily into two camps: the ethnographic and the formal. In the former category, Hine's 2006 study describing the creation of a mouse genome database, as well as Bietz and Lee's excellent 2009 ethnography of a metagenomics database's use and development are motivating touchstones for this work: many of our research questions were inspired by their work. However, we're interested in exploring database use, and its effects, at a longer time scale than traditional ethnography would necessarily allow. We are also interested exploring whether the phenomena these scholars describe continue to be found at that longer time scale. For instance, Hine finds that databases do not seem to fundamentally change how scientific work is done; we wonder if this continues to be the case for long-lived databases: databases that continue to be used two, five, ten and more years after their development.

On the formal end, MacKenzie's 2012 exploration of "how databases multiply" also seeks to answer research questions similar to ours, but from a mathematical perspective. MacKenzie frames the growing need for aggregation in and through databases via an exploration of "multiples" -- the intersection, division and making of sets as a form of world making -- relying heavily on Alain Badiou's "philosophical effort to articulate mathematics," specifically set theory, "as an ontology." Through his "mathematical orientation,"

MacKenzie seeks to show how databases "wrest inclusion from belonging" -- or make present social orderings through mathematical groupings. Though MacKenzie's insights are compelling, here we are more concerned with the day-to-day social interaction with sets: what it is for a non-mathematical human brain to adopt and maintain a set theoretic style of "relational thinking" to do work over time?

Manovich (1999) similarly discusses databases as sets of objects or datums -- a way of representing the world as "as a list of items which [the database] refuses to order." Manovich's characterization of a database's contents is an idealization based on set theory: the database provides unordered access to an item or set of items, which are retrieved via a structured query, as opposed to devices like catalog ledgers, which organize information according to a chronological narrative.

We counter, though, that while databases may not impose a specific order, the humans entering data into said database leave an archeological ordering or narrative behind -- both explicitly and implicitly. Explicit "narratives" include fields that mark the date of a record's creation, creator and last modification. Implicit narratives, on the other hand, are found in subtle changes to the ways that the database is used over time: the process of creating an entry may subtly change (new entries may be more or less complete than old); data entry conventions may shift (new users may bring with them new shorthand); and new fields may be added to existing tables, which would remain empty for older records. Databases such as wikis, which preserve their full unaltered histories, contain both explicit and implicit narratives, and consequently support commentary on changes in usage and conventions over time. Thus, we argue that the idea that database records are atemporal in their ordering is misleading -- as those who actually use the database are typically perfectly well aware. All databases contain some trace of their history to a greater or lesser extent, and older databases obviously have more history to leave a trace. These implicit and explicit narratives make trace ethnography feasible (Geiger & Ribes, 2011).

We believe (as Baker and Bowker did before us, in a closing note to their 2007 work on information ecology) that applying Brand's considerations of how buildings "learn" to databases will help bridge ethnographic and formal approaches, and provide and interesting framework with which examine use over time. Brand's perspective brings some of the long term thinking that is so readily present in museum work, preservation work, yet is weirdly absent or only just nascent in CSCW or HCI (Voida, Harmon and Al-Ani 2011 a notable example of the nascent). Brand's considerations of change over time, evolution of structure, and alteration of existing structures for unanticipated use are all apt for studies of databases.

3 Case Study: The Universal Chalcidoidea Database

Our case study focuses on a taxonomic database describing an interesting, but somewhat obscure group of organisms: the chalcid wasps. Described as "gem-like inhabitants of the woodlands by most never seen nor dreamt of" (Girault, 1925), these parasitic wasps are beautiful, plentiful, but often miniscule, making them hard to collect and study. There are an estimated 500,000 chalcids wasps in existence, yet only 22,000 have been named and described (Noyes, 2003). The Universal Chalcidoidea Database (UCD) is one taxonomist's efforts to collect and make available all existing literature on these organisms and their nomenclature.

Like many natural history databases, the now electronic UCD was once made of paper: in this case a "taxonomic card catalog" – a database of a group of organisms' names, namers and name changes – "compiled and maintained until about 1969 by the 'indexing section' of the Department of Entomology at The Natural History Museum, London" ("History of the Database"). In the mid 1970s, Dr John Noyes, a specialist in wasps, joined the staff and began augmenting the taxonomic index with an exhaustive bibliography of related literature; though exact use metrics have not been published, the database's website describes resulting card catalog and bibliography as being "constantly in use" by researchers on the group. In 1991, Noyes began migrating the card catalog to an electronic database; in 1998 this database was made available on CD-ROM; and in 2002 it was migrated to the "Taxapad" data management system. In August

2011, Noyes began work with the INHS to integrate the UCD with a larger system of taxonomic databases: Species File Software, maintained by the Illinois Natural History Survey (INHS).

3.1 Methods and goals of migration (or, how we learned the database)

The UCD database was given to us in a number of formats, some more usable than others (e.g. databases in proprietary file formats that couldn't be read without their originating program; a folder containing a number of text files; and finally some SQL dumps that lacked relations). We decided to work with the SQL dumps, and try to reverse engineer the relations between tables. We also had Noyes' documentation for the database, initially written to aid student workers at the NHM with data input and database maintenance. These materials included an entity-relationship diagram, of sorts (Figure 1), as well as an extensive pdf describing the contents of each field in each table. We encountered some usual and expected rough spots with migration: despite the schema diagram and other documentation, we still needed to contact Noyes to clear up questions. The meaning of some table and field names were unclear -- some because of technical constraints (e.g. character limits to field names; field limits to tables) and some because of our inexperience with the database. In other cases, though it was because the actual structure of the database had been changed from its original design: Noyes' database had "learned," so to speak, to adapt to the changing conditions in his lab.

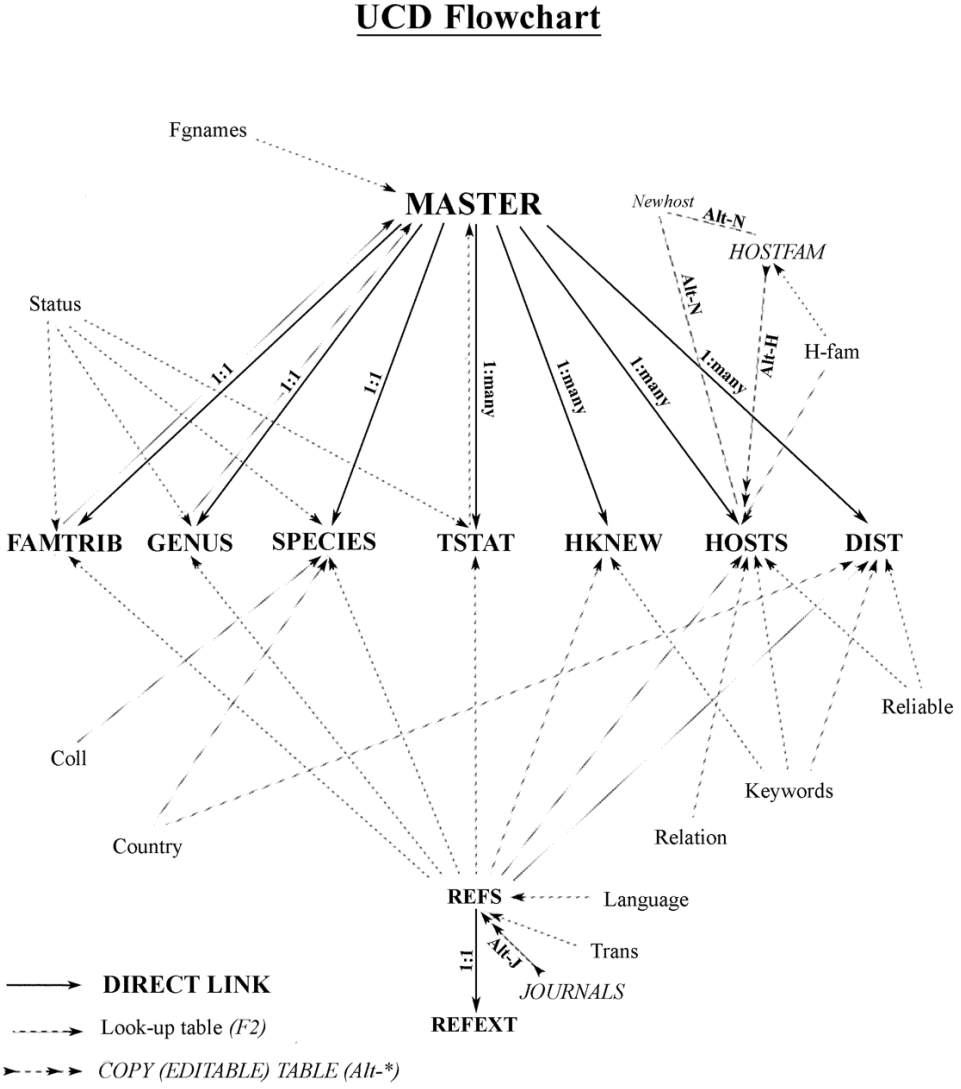

Figure 1: Noyes' "Flow Chart" describing his database's original schema

4 Discussion

4.1 How this database learned

In some ways, it's insufficient to only refer to the UCD as a database – it's also an extremely well-curated dataset. Noyes often had to rely on non-expert, unpaid volunteers for data entry, so he had to stringently checked all of their work before "accepting" it into the database. However, instead of adding additional fields into the primary set of tables, Noyes created proxy tables into which volunteers could enter their data. Noyes then would manually migrate these new records to the main set of tables. This double layering of tables isn't reflected in the database's schema (Figure 1) because it was a later addition to the database's structure; we only realized what was going on after encountering seeming duplicates or versions of the same table ("ref" and "newref") and contacting Noyes for clarification.

After more thoroughly comparing Noyes "Flowchart" and the sql dump we were given, we realized that Noyes' add-on construction had been quite extensive. Our database includes 34 tables, whereas Noyes' original schema only contains 22. Again, after consulting with Noyes, we learned that he had "ingested" several other datasets into his over the course of the UCD's lifespan, and had furthermore, begun using the

UCD for local data management of some related-but-separate projects, such as a table titled "crencyrt" which contains data from survey of Costa Rican *Encyrtidae* ranges otherwise unrelated to the rest of Noyes' data aggregation efforts.

As we implied through our literature review, much of the prior work in databases has been either ethnographic, and therefore difficult to generalize, or extremely mathematical -- only looking at database contents and change over time in terms of set theory and mathematical relationships. Here we want to show how sets and workplace practices affect each other over time, and Brand's framing allows us to do just that: to study the interplay between engineering, culture, and the day-to-day getting on with it. In the case of the UCD, the changes to this database particularly lend themselves to Brand's architectural metaphors: tables were "added on" like spare rooms to make room for an expanding "family" of users. Because Noyes so carefully curated his data, we did not find some of the quirks of long-term use that we have observed in our own prior work with databases, such as gradual change in the use of certain fields over time (the repurposing of a room, in Brand's rendering), or of shearing of large tables into smaller subsections.

5 Conclusion

To borrow further from Brand, we believe that it is useful to view a database not just as a completed product, but as something that is in the process of change; the process of database construction is not confined to the period of its original design. As with houses, older databases contain evidence of how they were built, how they have changed, and sometimes even why. In the UCD, we see evidence of design processes that allow for the safe handling possibly erroneous data entry, as well as repurposing the database for a subproject. We were able to take advantage of various supplementary sources of evidence to inform this analysis in addition to the database itself.

We believe that studies like this can give us a richer understanding of how longer lived databases subtly change over time, thereby informing not only their initial design and management, but their long-term maintenance and preservation as well. It is hardly controversial to argue that databases should last – but we want to ask, how do they survive, really – particularly now that our longest lived databases (those found in memory institutions like libraries and museums) are becoming increasingly electronic, and thus increasingly at risk for being treated as wholly mathematical entities – not the complex human created artifacts that they in fact are. Many present preservation techniques further put them at risk for being preserved as pristine, atemporal entities; rather like historic homes that have never changed from the moment that they were built, databases in preserved in this manner will be seen but not used. Understanding how databases learn will help us understand the features that contribute to databases' long-term usefulness and usability over years, decades, and even centuries.

6 References

Baker, K. S., & Bowker, G. C. (2007). Information ecology: open system environment for data, memories, and knowing. *Journal of Intelligent Information Systems*, *29*(1), 127–144. doi:10.1007/s10844-006-0035-7

Bietz, M., & Lee, C. (2009). Collaboration in metagenomics: Sequence databases and the organization of scientific work. *ECSCW 2009*, (September), 7–11. Retrieved from http://www.springerlink.com/index/t7124470143464r9.pdf

Brand, S. (1995). *How Buildings Learn: What Happens After Their Built.* Penguin Books.

Girault, A.A. (1925). "Some Gem Like Inhabitants of the Woodlands by Most Never Seen Nor Dreamt Of." *The Literature of Platygastroidea.* Retrieved from http://plazi.org:8080/dspace/handle/10199/15794

Geiger, R. S., & Ribes, D. (2011). Trace Ethnography: Following Coordination through Documentary Practices. *2011 44th Hawaii International Conference on System Sciences*, 1–10. doi:10.1109/HICSS.2011.455

Hine, C. (2006). Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work. *Social Studies of Science*, *36*(2), 269–298. doi:10.1177/0306312706054047

"History of the database." Universal Chalcidoidea Database. Retrieved from http://www.nhm.ac.uk/research-curation/research/projects/chalcidoids/database/. 10 September 2013.

Mackenzie, A. (2012). More parts than elements: how databases multiply. *Environment and Planning D: Society and Space*, *30*(2), 335–350. doi:10.1068/d6710

Manovich, L. (1999). Database as Symbolic Form. *Convergence: The International Journal of Research into New Media Technologies*, *5*(2), 80–99. doi:10.1177/135485659900500206

Noyes, J.S. 2003. Universal Chalcidoidea Database. World Wide Web electronic publication. http://www.nhm.ac.uk/chalcidoids

Schuurman, N. (2008). Database Ethnographies Using Social Science Methodologies to Enhance Data Analysis and Interpretation. *Geography Compass*, *2*(5), 1529–1548. doi:10.1111/j.1749-8198.2008.00150.x

Voida, A., Harmon, E., & Al-Ani, B. (2011). Homebrew databases: Complexities of everyday information management in nonprofit organizations. In *Proceedings of the 2011 Annual Conference on Human Computer Interaction* (pp. 915–924). Retrieved from http://dl.acm.org/citation.cfm?id=1979078

7 Table of Figures