Taxonomic Work as Information Work: Design for Semantic Refactoring

Andrea K. Thomer
Michael B. Twidale
Jinlong Guo
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA
thomer2@illinois.edu.edu
twidale@illinois
jguo24@illinois

Matthew J. Yoder
Illinois Natural History Survey
University of Illinois at Urbana-Champaign
Champaign, IL 61820, USA
mjyoder@illinois.edu

CSCW '16, February 27 - March 02, 2016, San Francisco, CA, USA

Abstract
Taxonomy is the branch of science concerned with classifying organisms: drawing the line between cats and dogs, fish and fowl, animals and vegetables. Modern taxonomic work is built on a hundreds-year-old tradition of qualitative research and description. There are aspects of this work that illustrate the pervasiveness and difficulty of a particular kind of qualitative data wrangling, which we call *semantic refactoring*: the review, normalization, and re-engineering of semantic structures. Because taxonomic work is conducted over long time spans, the processes underlying semantic refactoring become more visible. An examination of taxonomic data practices may inform our understanding of how (and if) collections of qualitative data scale, particularly when collaboratively created.

Author Keywords
Scientific workflows; qualitative data; taxonomy; human-information interaction; biodiversity informatics; qualitative data; ontologies; classification

ACM Classification Keywords
J.3 [Life and Medical Sciences]: Biology and genetics; H.5.2 [Information Interfaces And Presentation]: User Interfaces: User-centered design

Introduction

Through the NSF-funded Transforming Taxonomic Interfaces project, we are studying (via semi-structured interviews, hackathons, and collaborative prototyping of interfaces) the day-to-day work of biological taxonomy; our goal is to better understand taxonomists' data practices and thereby build better interfaces and tools for them. Though taxonomy (the hundreds-year-old branch of science concerned with classifying life on our planet) represents a typical "small science" in some ways, the data it produces are quite big: there are an estimated 8.7 million species on this planet, all of which need names, descriptions, and delineations from another, so that we can better understand earth's biodiversity. However, unlike other big data, taxonomic data are fundamentally qualitative: they are the result of expert assessment, and presented primarily as words, images (photographs and sketches), and physical objects (type specimens) - not numbers.

Like other fields of biology, taxonomy has become increasingly computational over recent decades. However, because of the field's unique reliance on qualitative and textual data, taxonomic informatics has as much in common with the digital humanities and computational social sciences as it has with bioinformatics. Because taxonomic literature is extremely long-lived, and researchers regularly require decades-old data, there is a large body of legacy data in need of migration to modern formats through text mining [3, 8]. Thus there is a growing need for what are essentially robust qualitative coding standards - formalized anatomical ontologies (e.g. [9]) - that facilitate inter-coder consistency, yet which don't artificially make a closed world of an open system.

Prior work by Bowker has treated taxonomy as a "model organism" for scientific memory practices, record keeping, and classification [1, 2]. We believe there are further aspects of taxonomic work that will be of interest and value to the study of human centered data science (HCDS). For instance, the work of "scaling up" qualitative data analysis necessarily will require normalizing and integrating large amounts of qualitative data. Taxonomists have been considering how to do just that for hundreds of years. Further, the specific processes involved in qualitative data wrangling are often more visible in taxonomy than in other big (qualitative) data fields, because of taxonomy's unique scale and scope. Taxonomists spend years, sometimes decades, working on a description of a group of organisms. This broad temporal scale forces them to make explicit commonly tacit tasks, such as the criteria for classifying something one way as opposed to another - if only so that they can remember their work from one month to the next. This provides us as CSCW researchers an opportunity to examine in rare detail systems that are ordinarily fast moving or invisible.

Additionally, taxonomists are experts at integrating and working with longitudinal qualitative data. Not only does their own work take years to complete, but they also must interpret and integrate data from decades and centuries past. Taxonomists must not only migrate data from one format to another (e.g. paper to spreadsheets to databases), but sometimes from older languages to newer (e.g. Latin to English). This work has wide-ranging implications for HCDS, particularly as it relates to the use and development of qualitative data interoperability standards.

Here we briefly describe the information work (e.g. [5]) of taxonomic work, and explore the implications this work has for HCDS. We emphasize the need to consider *semantic refactoring* as a critical yet under supported task in CSCW. In doing so, we point to a need to unpack the traditional dichotomy between qualitative and quantitative data.

Taxon	5 fingered hand; claws	Fur; milk
Bear	X	X
Lion	X	X
Lizard	X	
Platypus	X	X
Tiger	X	X
Zebra		X

Figure 1: A very simple character matrix. Each "X" notes the presence of a "character" in an organism.

Taxonomic work, taxonomic data

The primary goal of taxonomic work is to create a taxonomic description: a semi-structured narrative describing what makes a group of organisms unique. This description may be written *de novo* (describing a new species altogether), or may be a revision of an existing description. Revisions are necessary when older descriptions are not clear, thorough, or otherwise complete enough to be useable in modern research. The following excerpt is from a description of a subfamily of wasps:

"Subfamily TETRACNEMINAE Mandibles bidentate; forewing without a filum spinosum; setae on basal cell of similar size to those beneath apex of venation).... FEMALE - gaster with last tergite more or less shield-shaped or triangular, its anterior margin almost straight, hardly curved" [4].

These descriptions are the result of a long, iterative process of reviewing prior work, collecting and sorting specimens, and identifying diagnostic "characters" (aspects of an organism's anatomy) that distinguish one group from another. For instance, the presence of fur distinguishes mammals from reptiles and birds; and in the example above, the presence of a mostly straight anterior edge of the last segment of an insect's "gaster" (tail end) distinguishes the *Tetracneminae* from other chalcid wasps. Descriptions are sometimes created along with a *character matrix*: a table that translates qualitative data into a more structured form. Different character states are binned into ranked categories or binarized according to presence or absence of characters (Figure 1).

In writing a description, taxonomists must describe anatomical characters as clearly and unambiguously as possible, so that others can use them to identify specimens, or future taxonomists can determine whether they have discovered a new species. However, taxonomic data's extremely long lifespan is in itself an obstacle: the descriptions can last longer than our common understanding of the terms used to write them. Though individual taxonomists try to use preexisting vocabularies when possible, often they are too arcane to be clarifying. For instance: the terms "tomentose", "floccose," "arachnoid," "hoary," and "lanate" all refer to kinds of fuzziness that might be seen on a leaf. Yet, their use may introduce more noise than signal to a modern reader. Colloquial terms like "woolly" - along with extensive photographs and sketches - may be ultimately more informative. Thus, while individual taxonomists strive for consistency, they still must make hundreds of *ad hoc* decisions about which standards to adhere to, and which to ignore.

Researchers have increasingly turned to formal ontologies as a way of clarifying their descriptions while also making them machine readable, in an attempt to "scale up" to big data for computational analysis. Yet, application of ontologies is just as practically challenging (if not more so) as application of controlled vocabularies. Researchers must still choose their terms carefully, and be mindful how they will be eventually be interpreted and used (by machines and man alike).

Semantic refactoring

We might think of taxonomic description as a process of semantic engineering, and the interpretation or revision of descriptions as semantic refactoring, similar to software refactoring: the piecemeal practice of making semantic structures clearer and more efficient. This work requires on-going assessment of individual words' and concepts' fitness-for-use in a description and knowledge base. The refactorer must first understand her stakeholders - who (or what) she is creating a system of terms for, and why - and must review and revise the existing body of terms, and the

relationships between those terms, to serve her stakeholders' needs.

Semantic refactoring is critical to supporting qualitative data interoperability, particularly in for long-lived, collaboratively created datasets. It is akin to the process of ensuring intercoder reliability, but over the course of many years and projects. Taxonomists provide us with an excellent case study for this task, as they have been on the forefront of qualitative data standards development and collaborative dataset creation since Linnaeus published his *Systemae Naturae* in 1735. Data standards in taxonomy are simultaneously rigid and flexible: Linneaus' Latin-based system of binomial nomenclature has remained in place despite challengers [6], yet the specific terms used in descriptions have been slowly changing. We believe this wrangling and refactoring process is a kind of articulation work common to qualitative research (particularly that which is gathered through qualitative coding, surveys, or other methods that "bin" amorphous phenomena into categories or classifications), yet is under-supported by current information systems and interfaces. The fields of CSCW and HCDS should seek to understand and design for this work.

Beyond quantitative and qualitative data

Studying taxonomists has forced us to reconsider binary distinctions between the terms "quantitative" and "qualitative" as applied to research, data, and methods. Even quantitative values in taxonomy are derived through qualitative processes in their collection or determination; something as seemingly simple as counting the number of legs on a creature requires first deciding what constitutes a leg (for instance, if fish don't have legs, but frogs do, what does *Tiktaalik roseae* [7] have? And how many?). We believe this points to a potential area for future study: what data types exist between qualitative and quantitative? And what do we lose when qualitative data are treated as if they are naively or natively quantitative? Natural language processing is certainly computational - but not necessarily quantitative. We need to examine the relationship between quantitative and qualitative data if we want to make progress in non-numerical computability.

Implications for CSCW

We find that an analysis of the information work of taxonomic work uncovers a range of issues that are likely to have broader implications for HCDS and CSCW. Taxonomy is an exemplar of a long-lived collaborative discipline; understanding how taxonomists work together to create qualitative datasets over great geographic and temporal distances will inform similar fields and existing initiatives in computational disciplines. Taxonomic work forces us to reconsider simple binary distinctions between qualitative and quantitative data, methods, and analysis that we believe have implications for the "scaling up" of big qualitative data. We have further argued that the process of semantic refactoring is a kind of articulation work particular to qualitative research, and one that would benefit from further study from a CSCW perspective.

Acknowledgements

This research was funded through NSF Grant 1356515.

REFERENCES

1. Geoffrey C Bowker. 1999. The game of the name: Nomenclatural instability in the history of botanical informatics. In *Proceedings Of The 1998 Conference On The History And Heritage Of Science Information Systems*. 74–83.
2. Geoffrey C. Bowker. 2000. Biodiversity Datadiversity. *Social Studies Of Science* 30, 5 (oct 2000), 643–683.

DOI:
http://dx.doi.org/10.1177/030631200030005001

3. Hong Cui. 2012. CharaParser for fine-grained semantic annotation of organism morphological descriptions. *Journal of the American Society for Information Science and Technology* 63, 4 (apr 2012), 738–754. DOI:http://dx.doi.org/10.1002/asi.22618
4. John S Noyes. 1988. Encyrtidae (Insecta: Hymenoptera). *Fauna of New Zealand* 13 (1988).
5. Carole L Palmer, Melissa H. Cragin, and Timothy P. Hogan. 2007. Weak Information Work in scientific discovery. *Information Processing & Management* 43 (2007), 808–820. DOI: http://dx.doi.org/10.1016/j.ipm.2006.06.003
6. Randall T Schuh. 2003. The Linnaean system and its 250-year persistence. *The Botanical Review* 69, 1 (2003), 59–78.
7. Neil H. Shubin, Edward B. Daeschler, and Farish A. Jenkins. 2006. The pectoral fin of Tiktaalik roseae and the origin of the tetrapod limb. *Nature* 440, 7085 (April 2006), 764–771. DOI: http://dx.doi.org/10.1038/nature04637
8. Anne E Thessen, Hong Cui, and Dmitry Mozzherin. 2012. Applications of natural language processing in biodiversity science. *Advances in bioinformatics* 2012 (jan 2012), 391574. DOI: http://dx.doi.org/10.1155/2012/391574
9. Matthew J. Yoder, István Mikó, Katja C. Seltmann, Matthew A. Bertone, and Andrew R. Deans. 2010. A Gross Anatomy Ontology for Hymenoptera. *PLoS ONE* 5, 12 (dec 2010), e15991. DOI: http://dx.doi.org/10.1371/journal.pone.0015991