Peter M. Logan, Emeritus Professor of English, Temple University
Victorian Data Conference, University of Virginia
11/14/19

"Transforming Nineteenth-Century Knowledge"

The Nineteenth-Century Knowledge Project is creating a digital data set of nineteenth-century knowledge in order to study how knowledge changes over time. This is "official" or dominant knowledge, and I'll explain why in a moment. We use historic editions of the *Encyclopedia Britannica* as a proxy for what counted as knowledge in the past. Britannica was the most authoritative general reference work of the nineteenth and much of the 20th century. First published in 1771, it continues in publication today and is the only encyclopedia in any language to survive that 250-year period. It has long been used by researchers to document changes in individual concepts over time, since it provides evidence for when a concept could be called "widely accepted." But this data has much more to tell us than what happened to individual concepts. Britannica's continuity gives us a unique opportunity to explore the broader question of how the structure of knowledge changes, by comparing different editions and identifying patterns in its transformation. And that is the goal of the 19th-C Knowledge Project.

There have been 15 editions of Britannica since 1771. Some were reissues with a few new articles, but four relevant to C19 were major editions with new versions of most entries. We based our corpus on these four, spanning the French Revolution to WWI.

1. *3rd ed. 18 vols., 1788-1797.*
2. *7th ed. 21 vols. 1830-1842.*
3. *9th ed. 25 vols. 1875-1889.*
4. *11th ed. 29 vols. 1910-1911.*

Altogether, these 93 volumes include about 100,000 entries totaling 125 million words. All of it has to be scanned to create standards-compliant TEI files with an error rate of <1%. Thanks to an award from the NEH, we will have that work completed in Spring 2020.

Britannica was a *comprehensive* reference source, unlike domain specific encyclopedias, like the *Domestic Encyclopedia* (1802), which focused on England and agriculture, or the self-explanatory *Brewer's Dictionary of Phrase and Fable* (1870). We know Britannica represented *all* knowledge, too. It even tells us so. They did take comprehensiveness seriously enough to recruit major figures from every field to explain the newest developments (Kogan 31 ff). Britannica paid contributors well, so they could get the biggest names to write in the field—Scott wrote on the Romance, Drama, and Chivalry. A remarkable list of Victorian intellectuals. But … wait; where are all the women? The first articles by women were not published until 1889, 118 years after the first edition. Mrs. Humphry Ward had a small critical biography of the Renaissance poet, dramatist, and courtier John Lyly … another male author.

The absence of women authors tells us little about Britannica, actually, but it illustrates how official knowledge works: As social beliefs change, so do the culture's ideas about what counts as "knowledge." Nineteenth-century social beliefs determined the gendered selection of authors. They also determined the selection and length of entries in the encyclopedia, where length indicates importance. There were few topics specific to women, as we would expect from a society that discounted their contributions. Articles on India and Africa reflected the colonizer's perspective and represented indigenous people in ways that were overtly racist and imperialist. There were no articles by writers of color in any of these editions.

This historical collection is not what we would call "knowledge" today, and that is exactly the point. Knowledge is socially constructed, and it changes over time as social beliefs

change. This is not a new idea. Karl Mannheim wrote about it in 1936: “The principal thesis of the sociology of knowledge is that there are modes of thought which cannot be adequately understood as long as their social origins are obscured” (Mannheim 2). Britannica was received as the most authoritative reference source in English because it faithfully represented the idea of knowledge to Victorians, as they imagined it. It was an elitist system, racist, and just *wrong,* but thanks to Britannica’s attempts at comprehensiveness, we now have a tailor-made data set for identifying the assumptions that governed decisions about what counted as knowledge in the nineteenth century, what did not, and why.

* * *

How exactly are we going to analyze this data, to expand and contribute to our knowledge about knowledge? Our first priority is to use a comparative method by subdividing the corpus and comparing measurable quantities, like how linguistic complexity in the different knowledge sectors compare. But what exactly are the sectors? Should we use the divisions of knowledge as currently understood? Or the divisions as they were thought at the time? I’d like to compare the humanities articles to those in science, for example, but the humanities did not exist as an accepted concept until the second half of the nineteenth century, and the concept of “letters” that preceded it included the sciences. In the rest of this talk, I want to explain what we are doing as we build the data set, and show you why taking knowledge as the object of study raises epistemological questions that require us to develop an unusual analytical technique tailored to the material.

As I mentioned, the complete data set is 125 million words. Analyzing a corpus of this size depends on classifying each entry and adding authoritative subject headings to every one of the 100,000 entries. The path from data to knowledge runs through the valley of metadata. This

is a topic too often overlooked or ignored outside of information science, or imagined as transparent and obvious—a kind of "black box" that makes library and internet searches happen, for example. But we ignore it to our peril, as I discovered in the course of trying to develop knowledge about knowledge.

We cannot of course classify entries by hand—that is the gold standard, but no DH projects that I know of have the resources to put a cataloger on staff, let alone the army of them we would need for 100,000 entries. Instead, we are working with an automated classifying program at the Metadata Resource Center at Drexel University, called "HIVE." It generates subject terms from a controlled vocabulary, like the LCSH. Automated metadata generators are still in their infancy, and relevancy rates are low; on average, they produce useful results about 40% of the time. But the Metadata Research Center is improving that baseline, so I'm fortunate to be working with them.

HIVE uses a classifying algorithm which looks at work frequency and word proximity, something like what LDA does in topic modeling, to identify what it thinks are the most important concepts in the entry. It then adds a second step: it compares these keywords to a controlled vocabulary—a specialist vocabulary of normalized terminology—and generates subject headings from the controlled vocabulary, rather than the most frequent words. Finally, it adds Linked Open Data references to online authority files for the chosen vocabulary. This last step optimizes our files for web search engines, making them highly discoverable.

We run entries in bulk through HIVE and add the metadata it generates to the entry's TEI header, before adding them to a database. By querying the database, we can segment the corpus into our sections of related entries—all entries about agriculture or those on engineering, for example. Using controlled vocabularies, rather than the words of the entry itself, is critical. For

example, if one author discusses “cultivated forests” and another mentions “artificial forests,” a simple word frequency test will treat them as different concepts. HIVE recognizes them as different terms for the same thing and uses the correct term from the controlled vocabulary.

I began this project thinking we would use the current Library of Congress Subject Headings to generate consistent subject metadata. Simple! While adding subject metadata from LCSH would benefit many DH projects, it poses a special problem when exploring the historical construction of knowledge. The central historiographic challenge we face is preserving the specificity of the system of knowledge that these older editions represent. Controlled vocabularies, including the magnificent LCSH, are themselves representations of a system of knowledge. Controlled vocabularies also reflect changes over time in how knowledge is organized and delimited. In other words, they are similar to the encyclopedias we are trying to analyze and they embody the same problems.

For example, when published in 1876, the first edition of the *Dewey Decimal Classification* (DDC) had exactly two subject terms for African Americans: #326, “Political Science—Slavery,” and #573, “Biology--Natural History of Man” (Dewey 16, 18). This leaves no room for books authored by African American writers or their contributions to science, social science, and the arts. This lacuna was not an oversight, any more than the absence of woman writers in Britannica was an oversight. Dewey mirrored the same racial stereotyping present in the nineteenth-century Encyclopedia editions, where people of color were present as races but absent as individuals. Both the controlled vocabularies and encyclopedias were products of similar social systems, and both embody similar perspectives in representing knowledge as a structured whole.

The current LCSH is of course far different from the 1876 DDC and more sensitive to problems of this sort. But that's not the point. There are no value-neutral controlled vocabularies, and the LCSH is simply a better reflection of our concept of knowledge than the first Dewey is. It continues to evolve, and eventually some beleaguered scholar will look back on its 2019 instantiation and think "WTF?" By using the current LCSH to generate subject headings for these historical encyclopedias, we transform their historical knowledge structure into our current mode of thinking. But this erases the older system instead of illuminating it. It muddies the waters precisely when we need to see most clearly the difference between our present assumptions about knowledge and the older ones that constitute our object of study.

We are creating digital versions of older vocabularies and adding them to HIVE to generate metadata about knowledge using vocabularies that retain its older structure. Digitizing these older vocabularies is both an intellectual and a technical challenge. First, none of them exist in usable form. This means we have to OCR text images. We then convert that text into the Rich Data Format standard defined by the W3C for interoperability in the semantic web, SKOS (Simple Knowledge Organization System).[1] Finally, they have to be mounted in HIVE.

But what is the most appropriate vocabulary for each edition. Only one of the four, the third (1797), tells us which vocabulary to use, and it provides a dramatic illustration of how much its vision of knowledge differs from our own. It cites the taxonomy that Ephraim Chambers devised for his *Cyclopaedia* in 1728. In the Preface to that work, Chambers notes the complexity of organizing knowledge into discrete topics. He wrote that the greatest difficulty "lay in the Form, in the Order, and the Oeconomy of the Work," so that the whole would not become "a confused Heap of incongruous Parts" (i). "Former Lexicographers have not attempted

[1] https://www.w3.org/2004/02/skos/

anything like a Structure in the Works," he explained, since they did not cross reference entries with related ones. So he invented his own taxonomy by considering all areas of knowledge "as so many Parts of some greater Whole" (i). His system is fundamentally Platonic, but also reflects early Enlightenment ideals and its emphasis on reason. He subdivides all knowledge into two great categories. "Natural" means the ideal or theoretical, in line with the Platonic Ideal. "Artificial" includes everything that is made or represented. Ethics, law, and science are thus Natural, including the new notion of innate human rights. Politics, jurisprudence, and husbandry are Artificial, because they concern specific implementations of ethics, Law, and science. This organization of knowledge is foreign to us today, and yet it formed the conceptual basis for the first editions of Britannica. And so we are using Chambers's taxonomy to generate subject terms for entries from that edition, in order to preserve its historical specificity as a representation of knowledge.

For the others, we will use the most prominent controlled vocabularies near the time of each edition's publication. Thus, for the 1911 edition, we are going to rely on the first edition of the Library of Congress Subject Headings, published one year earlier. And we assume that It is a reasonable historical representation of the Britannica's implicit model of knowledge organization. Earlier, we have the same first edition of the Dewey Decimal Classification system, from in 1876, which we will try on Britannica's seventh edition of 1889.

Currently, we have completed the SKOS version of Chambers and added it to HIVE; [briefly explain the interface.] We also have the first volume of the two-volume 1910 LCSH added to HIVE and are working on the second. Our first tests with Chambers has allowed us to test one theory: that historical vocabularies will produce more relevant results than contemporary ones. For one thing, using the current LCSH to label historical entries can produce anachronistic

errors. We selected thirty sample entries from the third edition and ran them through both Chambers and the current LCSH. We then tested whether or not they were relevant to the entries. In the test, the Chambers vocabulary performed better than LCSH, increasing relevant returns from 51.88% to 60%. That doesn't seem like a lot, but it is an excellent result for an automated metadata generator, where every percentage increase matters. Since no one has ever tried the technique before, it also gives us some early data to justify pursuing this approach. And it suggests to me that nineteenth-century DH scholars might want to consider working with historical vocabularies to index their own data, rather than contemporary ones.

Chambers, Edwin. *Cyclopaedia; or, an Universal Dictionary of Arts and Sciences*. Vol. 1. 2 vols. London, 1728. *Internet Archive*. Web.

Daston, Lorraine, and Peter Galison. *Objectivity*. New York: Zone Books, 2007. Print.

Dewey, Melvil. *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*. Amherst, MA: Dewey, 1876. *Hathi Trust*. Web.

Kogan, Herman. *The Great EB: The Story of the Encyclopaedia Britannica*. Chicago: University of Chicago Press, 1958. Print.