Using Collections and Worksets in Large-Scale Corpora: Preliminary Findings from the Workset Creation for Scholarly Analysis Project

Harriett E. Green[1], Katrina Fenlon[1], Megan Senseney[1], Sayan Bhattacharyya[1], Craig Willis[1], Peter Organisciak[1], J. Stephen Downie[1], Timothy Cole[1] and Beth Plale[2]

[1] University of Illinois at Urbana-Champaign

[2] Indiana University

Abstract

Scholars from numerous disciplines rely on collections of texts to support research activities. On this diverse and interdisciplinary frontier of digital scholarship, libraries and information institutions must 1) prepare to support research using large collections of digitized texts, and 2) understand the different methods of analysis being applied to the collections of digitized text across disciplines. The HathiTrust Research Center's Workset Creation for Scholarly Analysis (WCSA) project conducted a series of focus groups and interviews to analyze and understand the scholarly practices of researchers that use large-scale, digital text corpora. This poster presents preliminary findings from that study, which offers early insights into user requirements for scholarly research with textual corpora.

Keywords: usability, digital collections, metadata

Citation: Green, H. E., Fenlon, K., Senseney, M., Bhattacharyya, S., Willis, C., Organisciak, P., Plale, B. (2014). Using Collections and Worksets in Large-Scale Corpora: Preliminary Findings from the Workset Creation for Scholarly Analysis Project. In *iConference 2014 Proceedings* (p. 1077–1083). doi:10.9776/14386

Acknowledgements: This work was supported by a generous grant from the Andrew W. Mellon Foundation. Additional support for conducting focus groups and interviews was provided by the Center for Informatics Research in Science and Scholarship at the University of Illinois' Graduate School of Library and Information Science. The HathiTrust Research Center is a collaborative initiative between the University of Illinois at Urbana-Champaign and Indiana University.

Contact: green19@illinois.edu, kfenlon2@illinois.edu, mfsense2@illinois.edu, sayan@illinois.edu, willis8@illinois.edu, organis2@illinois.edu, jdownie@illinois.edu, t-cole3@illinois.edu, plale@cs.indiana.edu

1 Introduction

To answer research questions about topics ranging from literary form to language and culture, humanities researchers may work with large numbers of complete volumes or smaller, hand-selected sets. While some researchers analyze the base texts, others interpret features derived from them.

Libraries and cultural-memory institutions must prepare to support research using large collections of digitized texts for analysis, or corpora, and need to understand the different methods of analysis applied to corpora across disciplines. The HathiTrust Research Center's Workset Creation for Scholarly Analysis: Prototyping Project (WCSA) conducted a series of focus groups and interviews to understand the scholarly practices of researchers using large-scale, digitized text corpora.

The HathiTrust (HT), a repository of over 10 million volumes (3 billion pages) of text, serves as a type of corpus: an expansive aggregation of distributed sources from which related sources may be concentrated by researchers into densely thematic bodies of evidence.[1] This aggregation consists of not just its primary constituents (books), but also the bibliographic metadata, and even intra-book content, such as formal sections, captioned images, maps and charts, and indexes. The HathiTrust Research Center (HTRC) is the research branch of the HathiTrust.[2] The HTRC offers a suite of tools and services, which enable computational access to the HT corpus. From digitized library collections in HT, scholars select subsets for

[1] http://www.hathitrust.org/htrc/

[2] http://www.hathitrust.org/

computational analysis according to their particular research objectives. We refer to these subsets, along with associated, external data sources, as "worksets." Worksets are a type of machine-actionable, referential research collection. User requirements for workset creation grow increasingly sophisticated and complex as humanities scholarship becomes more interdisciplinary and more digitally-oriented over time.

HTRC holds transformative promise for humanities scholarship: it seeks to enable scholars to sift through a massive corpus and to construct therefrom precise worksets for investigation. How scholars use collections and worksets remains a central research question in this initiative. Under the auspices of the HTRC, the WCSA team conducted a series of focus groups and interviews investigating how to facilitate scholarly selection of digital research materials.

WCSA is a two-year effort, funded by the Andrew W. Mellon Foundation, which aims to engage scholars in designing tools for exploration, location, and analytic grouping of materials so that they can routinely conduct computational scholarship. The three major goals of the WCSA project are to 1) enrich the metadata in the HT corpus, 2) improve access and discovery through reference-able metadata, and 3) formalize the notion of collections and worksets in the context of the HTRC. This study gathers qualitative data on scholarly practices with text corpora to inform the development of tools and services for HTRC.

2 Background

The use of digitized, primary source materials is growing in value and prominence among humanities scholars (Brogan, 2006; Palmer, 2005). In addition, the act of bringing together related information from various kinds of collections is essential to their research processes (Warwick, et al., 2008; Sukovic, 2008; Sukovic, 2011). In the course of their work, researchers create their own "digital aggregations of primary sources and related materials that support research on a theme" (Palmer, 2004).

Scholars rely on collections of texts to support research activities across numerous disciplines, ranging from physics and public health to English and computer science (Underwood, 2013; Argamon, et al., 2009; Heuser & Le-Khac, 2012; Moretti, 2009; Petersen et al., 2012). In certain domains, scholars create personal, digital carrels, gathering subsets of texts amenable to in-depth analysis using advanced tools and services (Mueller, 2010). Research collections comprise a variety of media and formats, which together function as a coherent collection of interwoven content and context (Brockman, et al., 2001). Previously identified scholarly needs for conducting research with digital collections include the need for bibliographic and evaluative tools for building thematic, curated sub-collections, infrastructure for ensuring sustainable and well-prioritized digitization of materials, and digital collections that cover a breadth of temporal and geographic areas (Palmer, et al., 2010; Proffitt, et al., 2008; Meyers 2010, 2011; Sinn, 2012).

Scholars also play a critical role in shaping how librarians and information scientists formalize collections to support research activities. A 2010 Council on Library and Information Resources (CLIR) report warned:

> While a greater reliance and dependency on digital resources is inevitable, the quality of the data and their organization and accessibility in service to teaching and scholarship are major concerns. Without the guiding voice of scholars, the tremendous effort now being devoted to digitizing our cultural heritage could in fact impede, not facilitate, future research. (CLIR, 2010).

In 2011, the Center for Informatics Research in Science and Scholarship at UIUC surveyed digital humanities scholars who were awarded Google Digital Humanities Awards and given large-scale text corpora from Google Books for their research projects. Among the major challenges and areas of need identified in the study's findings were 1) identifying and retrieving materials and 2) identifying characteristics of textual content. The authors noted:

> Researchers do not necessarily need huge sets of data to do interesting work, but the implication is that they do need flexible data delivery services that can deliver different kinds of data in different

> formats based on different searches for different kinds of research at different times. (Varvel & Thomer, 2011)

Developing such flexible services requires ongoing inquiry into the research practices of specific disciplines working with these sources, including investigation into the types of research questions posed by scholars and the types of analytical methods employed.

3 Methods

This study addresses the research question: How do researchers, especially humanities scholars, use collections in the course of their research, particularly in the context of textual corpora? The WCSA team collected data through semi-structured focus groups and interviews, which targeted researchers in the humanities and others working with digital collections.

Participants were asked about how they identify, select, and obtain access to texts for inclusion in analysis; transformation and pre-processing steps; units of analysis (works, manifestations, pages, n-grams OCR, images, etc.); methods of analysis; problems encountered in obtaining text corpora and materials not currently existing in digital form; and challenges to working with these digital collections (e.g., OCR quality, duplication).

Focus groups and interviews were conducted at the Digital Humanities 2013 conference, the 2013 Joint Conference on Digital Libraries, and the 2013 HTRC UnCamp. Thirteen individuals participated in the focus groups and five scholars were interviewed, for a total of eighteen participants in the study thus far.

Focus group and interview recordings were transcribed, and transcriptions are being manually coded to identify emergent themes. Each transcription is coded multiple times to ensure inter-coder reliability. Further content analysis is ongoing.

4 Preliminary Results and Discussion

Participants included junior and senior faculty at liberal arts colleges and universities, computer programmers, librarians, data scientists, academic technologists, and graduate students. Scholars were specialists in English literature, classics, linguistics, library and information science, and history. Participants were affiliated with academic institutions located around the world, including Great Britain, Singapore, Germany, France, and different regions of the United States.

A set of key themes have emerged from preliminary analysis. The following three examples illustrate the roles of collections; the need to implement granular, actionable units of analysis; and the importance of expert-enriched, shareable metadata.

1) Researchers consider the processes of collecting and workset-building to be basic scholarly activities. Researchers collect on the bases of diverse criteria, but aim for exhaustiveness within defined analytic constraints: for example, complete representation of a genre over some period of time, complete representation of the works by a demographic, or a complete lexicon of some language, in print, for a certain time period (Figure 1).

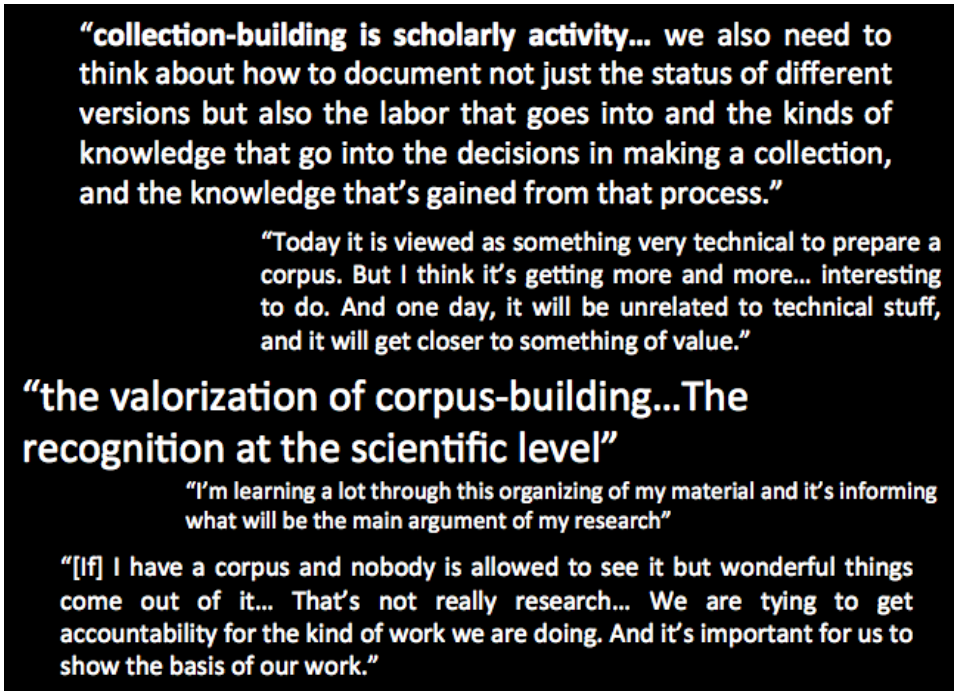

Figure 1: Selected focus group and interview excerpts on collection- and workset-building.

Some noted that, although the process of collecting and workset-building involves intellectually rigorous labor involving careful and refined analysis, its value may go unacknowledged by the scholarly community. There may be an interesting analogy to be made here with the often under-recognized intellectual work invested in the preparation of scholarly, edited collections of texts (Fraistat, 2012). This is an activity that is increasingly likely to become a species of workset creation, as the workflows (for preparing both print and scholarly editions) become more and more digitized over time.

2) Researchers desire that collections, worksets, texts, and other objects of analysis be highly divisible, and that resultant pieces be identifiable, movable, and readily associable with highly granular metadata--what Mueller calls "re-diggable and multiply recombinable data" (Mueller, 2012). Participants described a range of targets for analysis: full authorial *oeuvres*, individual novels, pages and page images, word tokens, n-grams (possibly tagged with parts of speech), poems within books, notions or themes, characters, encoded TEI elements, and lexicons, and more. They want to move subsets of worksets, or different logical or syntactic pieces, of their data between tools, collections, processes, formats, and standards, and to track them throughout (Figure 2).

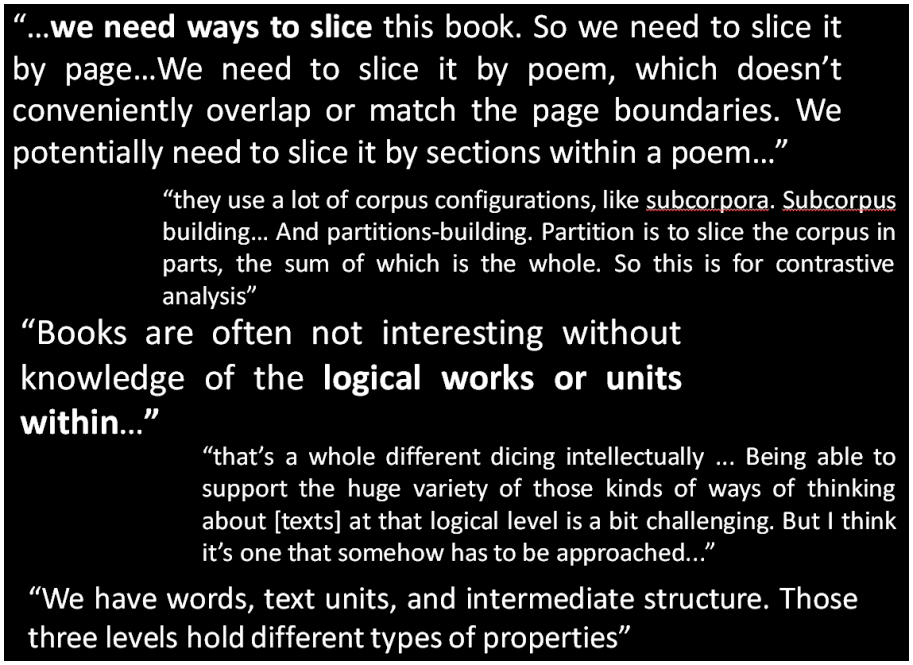

Figure 2. Selected focus group and interview excerpts on divisibility and objects of analysis.

The HathiTrust corpus is arguably better poised to support such a service than other existing repositories such as the Google Books corpus, for two reasons: first, because the HathiTrust, with its strong roots in research libraries, is more oriented towards serving the scholarly community than Google, the latter being more oriented towards serving a wider constituency in which the general public is a more significant component; secondly, Google, with its roots in information retrieval, is oriented towards searching for and retrieving information globally (its motto is "to organize the world's information"), but not specifically oriented towards differentiation or sectioning of search spaces, which is what the idea of worksets is focused on. This emphasis on differentiated and sectioned spaces that distinguishes WCSA is particularly relevant to the humanities. Unlike the sciences, which seek to discover universal and immutable laws of nature, the humanities are typically concerned with contextual relationships, that is, the relationships that obtain between texts in relation to a specific, situated, context of inquiry. This is likely to continue to remain true when the methods applied to humanistic inquiry are digital, as noted by Matthew Jockers (Jockers, 2013). Another literary scholar, Andrew Piper, makes a similar point about the needs that are left unmet by, paradoxically, that very globality, and lack of fine-grained differentiability, and, in particular, the lack of personalized sectioning/slicing of the search space in the Google Books affordances (Piper, 2013).

3) Researchers critically need more and better metadata, beyond conventional bibliographic metadata, for multiple aspects of the scholarly research process—from precise retrieval of texts to defining units of analysis. Participants noted a common desire to share their expert-created or -enriched metadata more broadly, much as they would disseminate results of uniquely created analytic work. Participants also expressed interest in collaborative, curatorial work on texts themselves, such as to edit, encode, or enrich the outputs of digitization (Figure 3).

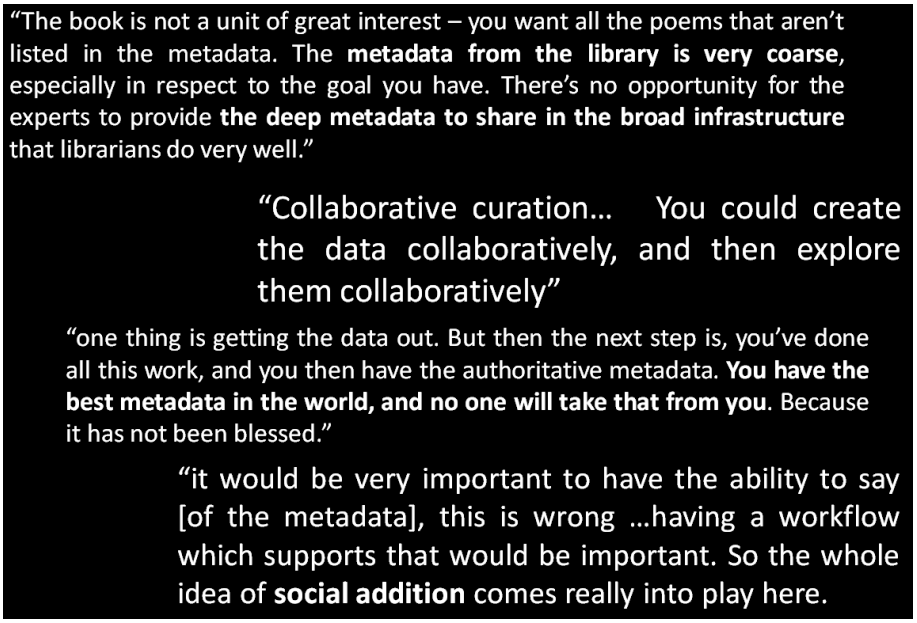

Figure 3: Selected focus group and interview excerpts on metadata enrichment and sharing.

5 Conclusion

Based on preliminary analysis, participants' responses indicate the need for formalized workset protocols that allow scholars to identify, select, and pull together subsets of texts within massive corpora. Ongoing data analysis will inform development of tools and services for HTRC, and best practices for other large-scale corpora. The study of user requirements for digital collections is critical to meeting the needs for rising levels of scholarly research with digital materials.

6 References

Argamon, S., Cooney, C., Horton, R., Olsen, M., Stein, S. & Voyer, R. (2009). Gender, race, and nationality in black drama, 1950-2006: Mining differences in language use in authors and their characters. *Digital Humanities Quarterly,* 3 (2). Retrieved from http://www.digitalhumanities.org/dhq/vol/3/2/000043/000043.html

Brockman, W. S., Neumann, L., Palmer, C. L., & Tidline, T. J. (2001*). Scholarly work in the humanities and the evolving information environment*. Washington, D.C.: Digital Library Federation, Council on Library and Information Resources.

Brogan, M. (2006). *Contexts and Contributions: Building the Distributed Library*. Digital Library Federation/Council on Library and Information Resources. Retrieved from http://www.diglib.org/pubs/dlf106

Council on Library and Information Resources (2010). *The idea of order: transforming research collections for 21st century scholarship*. Washington, D.C.: Council of Library and Information Resources.

Fraistat, N. (2012). Textual Addressability and the Future of Editing. *European Romantic Review*, 23(3), 329-333.

Heuser, R. & Le-Khac, L. (2012). *Stanford Literary Lab Pamphlet 4: A quantitative literary history of 2,958 Nineteenth Century British Novels: The Semantic Cohort Method* Palo Alto: Stanford Literary Lab.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Moretti, F. (2009). Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850). *Critical Inquiry*, 36 (1), 134-158.

Mueller, M. (2010). *Towards a Digital Carrel: A Report about Corpus Query Tools*. Retrieved from http://panini.northwestern.edu/mmueller/corpusquerytools.pdf

Mueller, M. Stanley Fish and the Digital Humanities. [Web blog post]. Retrieved from http://cscdc.northwestern.edu/blog/?p=332

Palmer, C. L. (2004). Thematic research collections. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.) *A Companion to Digital Humanities*. Oxford: Blackwell. Retrieved from http://www.digitalhumanities.org/companion/

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. *Journal of the American Society for Information Science and Technology*, 56(11), 1140-1153.

Petersen, A. M., Tenenbaum, J., Havlin, S., & Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports* 2, 313. doi:10.1038/srep00313

Piper, A. Beyond the e-Book: The New World of Electronic Reading. *World Literature Today*, 87(6), November 2013.

Sukovic, S. (2008). Convergent flows: Humanities scholars and their interactions with electronic texts. *Library Quarterly*, 78(3), 263-284.

Sukovic, S. (2011). E-Texts in Research Projects in the Humanities. In A. Woodsworth & W. D. Penniman (Eds.), *Advances in Librarianship* (131-202). Bingley, UK: Emerald Group Publishing.

Underwood, T. (2013). We don't already understand the broad outlines of literary history. [Web blog post]. Retrieved from http://tedunderwood.com/2013/02/08/we-dont-already-know-the-broad-outlines-of-literary-history/

Varvel, V. E. Jr., & Thomer, A. (2011). *Google Digital Humanities Awards recipient interviews report*. CIRSS Report No. HTRC1101. Champaign, IL: Center for Information Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Warwick, C., Terras, M., Huntington, P., & Pappa, N. (2008). If you build it will they come? The LAIRAH Study: Quantifying the use of online resources in the arts and humanities through statistical analysis of user log data. *Literary and Linguistic Computing*, 23(1), 85-102.

7 Table of Figures