Using Collections and Worksets in Large-Scale Corpora: Preliminary findings from the Workset Creation for Scholarly Analysis (WCSA) Prototyping Project

iConference 2014, March 4–7, 2014

Harriett E. Green, UIUC ; Katrina Fenlon, UIUC ; Megan Senseney, UIUC ; Sayan Bhattacharyya, UIUC ; Craig Willis, UIUC ; Peter Organisciak, UIUC ; J. Stephen Downie, UIUC ; Timothy Cole, UIUC ; Beth Plale, IU

Workset Creation for Scholarly Analysis: Prototyping Project

WCSA website: http://worksets.htrc.illinois.edu/worksets/

Project Objectives

- Enrich the metadata in the HT corpus,
- Augment string-based metadata with URIs to leverage discovery and sharing through external services, and
- Formalize the notion of collections and worksets in the context of the HTRC.

Research Question

Libraries and cultural-memory institutions must prepare to support research using large collections of digitized texts for analysis, corpora, and need to understand the different methods of analysis applied to corpora across disciplines. 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. We refer to these smaller sets, 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.

The WCSA Project team conducted a series of focus groups and interviews to address the following research question: How do researchers, especially humanities scholars, use collections in the course of their research, particularly in the context of textual corpora?

User Requirements Study

Members of the WCSA project team conducted a user requirements study in summer and fall 2013 to understand the scholarly practices of researchers using large-scale, digitized text corpora.

- Conducted semi-structured focus groups and interviews at the Digital Humanities 2013, Joint Conference of Digital Libraries 2013, and HathiTrust Research Center UnCamp 2013 conferences;
- Study contained 13 focus group participants and 5 individual interviews with humanities scholars, information scientists, digital humanities researchers, and librarians;

- Multiple rounds of qualitative coding using AtlasTI 7 for inter-coder reliability.

Early Findings

- 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.
- 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). 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.
- 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.

Next Steps

- Formalize workset model to allow researchers to identify, select, and pull together subsets of texts within massive corpora;
- Implement worksets in HathiTrust Research Center;
- Employ preliminary findings about user requirements to inform further tool-building prototyping projects to be awarded by WCSA.

HathiTrust Research Center

The HathiTrust Digital Library (http://hathitrust.org) is a repository of over 10 million volumes (3 billion pages) of text. The HathiTrust Research Center (HTRC) is the research branch of the HathiTrust. The HTRC offers a suite of tools and services, which enable computational access to the HathiTrust corpus. Learn more at: http://www.hathitrust.org/htrc/

Questions? Comments?

Harriett Green: green19@illinois.edu; @greenharr **Katrina Fenlon:** kfenlon@illinois.edu; @kfenlon

References

Mueller, M. (2012, February 8). Stanley Fish and the Digital Humanities. [Web blog post]. Retrieved from http://cscdc.northwestern.edu/blog/?p=332