Building worksets for scholarship by linking complementary corpora

Kevin Page
kevin.page@oerc.ox.ac.uk
University of Oxford, United Kingdom

Terhi Nurmikko-Fuller
terhi.nurmikko-fuller@anu.edu.au
University of Oxford, United Kingdom

Timothy Cole
t-cole3@illinois.edu
University of Illinois, United States of America

J. Stephen Downie
jdownie@illinois.edu
University of Illinois, United States of America

Background and General Motivation

The HathiTrust Digital Library

The HathiTrust Digital Library (HTDL) comprises digitized representations of 15.1 million volumes: approximately 7.47 million book titles, 418,216 serial titles, and 5.3 billion pages, across 460 languages. HTDL is best described as "a partnership of major research institutions and libraries working to ensure that the cultural record is preserved and accessible long into the future".

The HathiTrust Research Center (HTRC) develops software models, tools, and infrastructure to help digital humanities (DH) scholars conduct new computational analyses of works in the HTDL. For many scholars the size of the HTDL corpus is both attractive and daunting: many existing DH tools are designed for smaller collections, and many research inquiries are facilitated by more focused, homogeneous collections of texts (Gibbs and Owens, 2012).

Worksets

In many, if not most, DH research endeavours, performing an analytical task across the whole HTDL is neither practical nor productive (Kambatla et al., 2014). For example, a tool trained to identify genre attributes of 18th century English language prose fiction may not be applicable to 20th century French poetry. The first step is to identify the subset -- of works, editions, volumes, chapters, pages -- to set an initial investigative scope and, indeed, subsequent iterative refinements of a subset as research proceeds. In a corpus as large and complex as the HTDL, finding materials and then defining the sought after subset can be extraordinarily difficult.

HTRC has come to call collections of digital items brought together by a scholar for her analyses a "workset", created to help the researcher build, manipulate, iteratively define and compare their collections. Reflecting upon input and advice from the DH community, Jett (2015) defines a workset as a machine-actionable research collection realised as:

1. An aggregation of members (volumes, pages, etc.);
2. Metadata intrinsic to the workset's essential nature (e.g., creator, selection criteria);
3. Metadata intrinsic to digital architectures (i.e. creation date & number of members);
4. Metadata supportive of human interactions (i.e. title & description);
5. Derivative metadata from workset members (e.g. format(s), language(s), etc.); and,
6. Metadata concerning workset provenance (e.g. derived from, used by, etc.).

Broadly, item 1 identifies the actual data used in an analysis; whereas the remaining metadata items describe the workset itself, aiding workset management throughout the research cycle.

Cross-corpus worksets

As alluded above, numerous criteria can be used to select the constituents of a workset; and several technological implementations could, in theory, realise worksets. In researching the design and realisation of worksets and associated tooling, we are also mindful to remain grounded in their practical application and the needs of scholarly users. We have therefore undertaken our work through discipline-based scenarios in which we can explore the strengths and weaknesses of the HTDL viewed through the prism of worksets.

We report one such exploration here, questioning *whether (relatively) small, well explored, and well understood corpora can be superimposed over the HTDL to aid navigation and investigation of the much larger and superficially understood HTDL collection?*

From a system perspective, a cross-corpus workset requires exposing *compatible* metadata (items 2-6 above) from multiple collections, first used to align

common elements, and then to assemble worksets. We take a Linked Data approach and achieve compatibility through ontologies, which might initially be bibliographic (and derived from library records) but should be iteratively extensible into the domain of the subject of study.

Examples in early English print

Early English Books Online Text Creation Partnership (EEBO-TCP) is a partnership with ProQuest and over 150 libraries and universities, led by Michigan and Oxford, to generate highly accurate, fully-searchable texts tracing the history of English thought and learning from the first book printed in English in 1473 through to 1700. Between 2000-2009 EEBO-TCP Phase I converted 25,000 selected texts from the EEBO corpus into TEI-compliant, XML-encoded hand-transcribed texts, subsequently freely released in January 2015.

In the work reported here, we have conjoined EEBO-TCP with a HathiTrust subset consisting of all materials described in their metadata as being in English and published between 1470 and 1700.

To ensure a prototype which simultaneously explored the fit of scholars' needs to the technology and exercised the technical challenges outlined in the previous section, we undertook a 'complete circuit' through the datasets (Figure 1). We: (i) ran a consultative workshop to choose investigations which might form the basis of worksets; (ii) used these abstract worksets to identify concrete requirements for the conjoined metadata; (iii) generated metadata from both corpora according to these specifications; (iv) aligned elements from both datasets in an overlapping superset; (v) realised the worksets identified in (i) using this metadata.

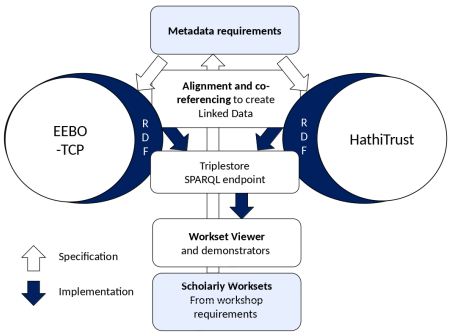

Figure 1. Overview of the metadata circuit leading to our cross-corpora workset

Motivating worksets

Following the workshop we identified the following workset selections; we describe their implementation in subsequent sections:

- Find all the works, appearing in both datasets, written by Richard Baxter.
- Find works in both datasets published in Oxford.
- Find works published outside of London (where the bulk were published).
- Find works from both datasets published outside of London in the mid-to late 1600s.
- Find all works in the two datasets for authors who have at least once published on the subject of "Political science".
- Find all works in these two datasets for authors who have at least once published works which are categorised as "biography".

Regarding the penultimate workset, it is of particular note that this returns results across both datasets, since our EEBO-TCP import did not contain genre or topic information; this association must be entirely inferred from the semantic links via the technology described below.

Implementation

Metadata requirements and ontology selection

Building on Nurmikko-Fuller et al. (2015) and Jett et al. (2016) we surveyed the addressable resources and the schema expressivity of ontologies that could parameterise these classes of workset. We identified parsable information structures in the EEBO-TCP TEI data, appropriate to the test worksets, and selected ontology terms to encode this EEBO-TCP metadata, ensuring compatibility (or at least, for our purposes, comparison) with RDF from the HathiTrust. The resultant ontology collection - the EEBO Ontology, or EEBOO - includes selections from MODS, Bibframe, and PROV, along with custom elements encoding additional structures (e.g. dates).

Creating EEBOO RDF and alignment with HTDL

Python scripts manipulated TEI P5 XML, then the Karma Data Integration Tool mapped EEBO-TCP data structures into the EEBOO ontology. Particular attention was paid to dates encoded within strings, an example of rich semi-structured data that can be extracted into structured RDF. Links to author records in

VIAF and the Library of Congress (LoC), and multimedia pages in the HTDL and 'JISC Digital Books' website, were generated and added. Finally, author names were aligned between the EEBOO and HTDL triples using a reconfiguration of the SALT tool (Weigl et al. 2016).

24,926 EEBO-TCP Phase 1 records were processed, with 22 distinct types of information in the headers, including 6 different ID types and 3 types of date (publication date of historical work, author associated historical date(s), XML publication/editing dates). EEBOO incorporates 7 of these datatypes, and extends into subcategories for author names and date types. EEBOO contains 713 unique places, 6,489 unique expressions of Person of which 3,588 have VIAF and LoC IDs.

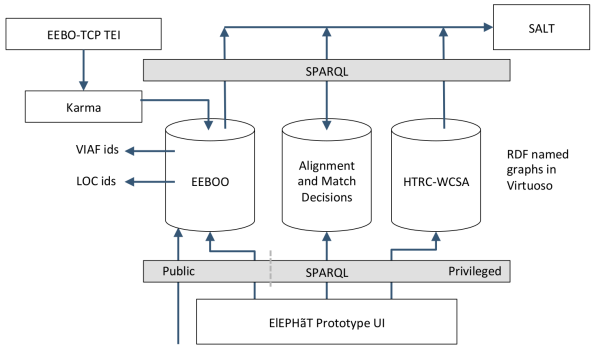

Figure 2. Architecture providing cross-corpus worksets for early English print

Workset construction and viewing

A Virtuoso triplestore (see also, the Virtuoso Github repository) stores the RDF data (totalling 1,137,502 triples) and provides a SPARQL query interface. Figure 2 shows the overall system architecture. The workset constructor user interface (Figure 3) allows the user to select parameters in a web interface which are, in the background, assembled into SPARQL queries used to create a workset. The interface automatically populates valid attributes that are themselves retrieved from the triplestore, using ontological terms having equivalent meaning across datasets. In combination, the generated triples and SPARQL queries are fully sufficient for expressing the motivating workset definitions described earlier.

The workset viewer (also Figure 3) then retrieves RDF workset contents, record metadata, data links, and multimedia links (to the Historic Books collection or the HTDL). Both web applications are written in Python, using the Flask framework, and both rely on the semantic information encoded in RDF and queried using SPARQL.

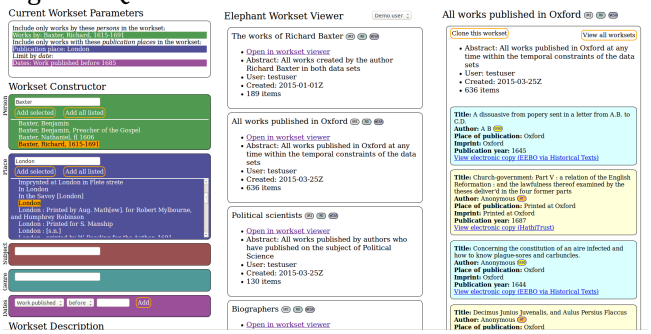

Figure 3. Prototype workset constructor and viewer (example worksets shown)

Conclusion and future work

We have demonstrated the general feasibility of cross-corpus worksets in bringing together HathiTrust content with specialised collections through a specific implementation for early English printed books linking the HathiTrust to EEBO-TCP. Using Linked Data, we see that metadata can be extended in a piecemeal or iterative fashion, potentially moving beyond traditional bibliographic metadata to include semantic structures emerging from scholarly investigation of the worksets themselves; and in doing so support academic motivations and requirements for workset creation.

Acknowledgements

We are grateful to our colleague Pip Willcox for her valuable input and organisation of scholars' workshop, and Jacob Jett for his workset ontology. This work was supported by the Andrew W. Mellon Foundation through the Workset Creation for Scholarly Analysis project award.

Bibliography

Gibbs, F., Owens, T. (2012). Building better digital humanities tools: Toward broader audiences and user-centered designs. Digital Humanities Quarterly 6(2). Accessible via: http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html

Jett, J. (2015). Modeling worksets in the HathiTrust Research Center: CIRSS Technical Report WCSA0715. University of Illinois at Urbana-Champaign. Available via: http://hdl.handle.net/2142/78149

Jett, J., Nurmikko-Fuller, T., Cole, T.W., Page, K.R., Downie, J.S. (2016). Enhancing scholarly use of digital

libraries: A comparative survey and review of bibliographic metadata ontologies. IEEE/ACM Joint Conference on Digital Libraries (JCDL) pp. 35-44, 2016.

Kambatla, K., Kollias, G., Kumar, V., Grama, A. (2014). Trends in big data analysis. Journal of Parallel & Distributed Computing 74(7), pp 2561-2573.

Nurmikko-Fuller, T., Page, K., Willcox, P., Jett, J. Maden, C., Cole, T., Fallaw, C., Senseney, M., Downie, J.S. (2015). Building Complex Research Collections in Digital Libraries: A Survey of Ontology Implications. ACM/IEEE Joint Conference on Digital Libraries (JCDL) p. 169-172, 2015.

Weigl, D. M., Lewis, D. L., Crawford, T., Knopke, I., Page, K. R. (2017, in press). On providing semantic alignment and unified access to music-library metadata. *International Journal on Digital Libraries.* Springer.