# **JETT — Prototyping**

# <2 images>

# **Prototyping a Workset Builder Using Semantic Technologies**

# Jett, J., Senseney, M., Maden, C., Fallaw, C. and Downie, J. S.

The HathiTrust Digital Library comprises more than 4.7 billion pages (12.6 million volumes) of digitized content. The HathiTrust Research Center (HTRC) is a collaborative research initiative led by the University of Illinois and Indiana University engaged in developing an array of tools to connect digital humanities researchers with materials of interest within the HathiTrust corpus. This poster discusses the activities of the Workset Creation for Scholarly Analysis (WCSA) project, an initiative of the HTRC. Part of the primary mission of the WCSA initiative is the development and evolution of worksets that include selected subsets of the HathiTrust corpus for use in computational analysis. To test how well semantic technologies fit the workset concept we have implemented a prototype RDF-based triple-store that allows scholars to directly engage with the metadata describing their worksets and the bibliographic entities.

A key component to this work is the development of an underlying formal conceptual model that effectively represents descriptive information about worksets, including provenance, curatorial intent, and other useful metadata, in a manner that facilitates the scholarly process of selecting, grouping, and citing research data collections. The prototype has been designed to (1) comply with standards established by the Linked Open Data and semantic web communities and (2) allow scholars the maximum amount of flexibility when gathering their research data collections together, permitting them to intermingle resources from external corpora with those contained within the HathiTrust Digital Library.

**Discussion**

As a majority (~66%) of the HathiTrust corpus remains under copyright, HTRC web services are being built primarily to provide “nonconsumptive” research. Under the nonconsumptive paradigm, the full contents of the copyright-restricted digitized books are never exposed to users, so scholars rely upon descriptive metadata about volumes within the corpora to assemble worksets. As depicted in the simplified workflow visualization in Figure 1 (below), scholars will then be able to submit their worksets to a number of analytics tools, both provided by the HTRC and developed by themselves. These processes will result in a number of data products that can be leveraged by the scholar in a number of ways, including as research materials that can be included in new worksets.

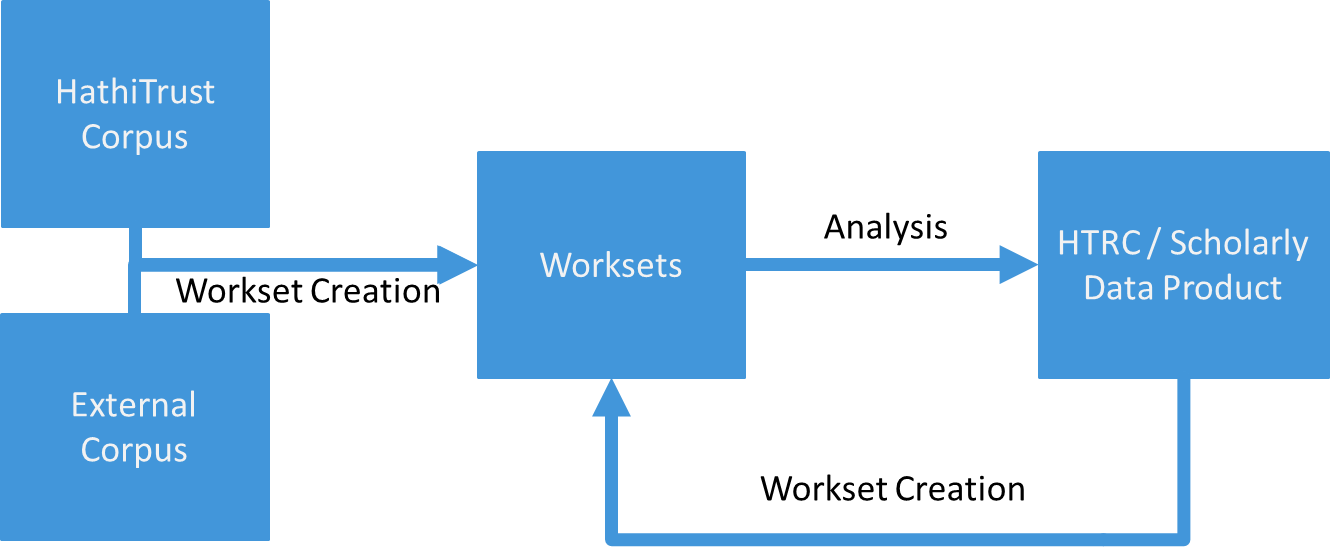


Figure 1. HTRC scholarly workflow.

Scholars require infrastructure that allows them to gather together masses of heterogeneous research materials (Varvel and Thomer, 2011), facilitates interoperability across datasets (Henry and Smith, 2010), and supports working with materials at arbitrary levels of granularity (Fenlon et al., 2014). These requirements have been a driving force in the development of our prototype and its underlying conceptual and data models. We built our prototype RDF-based workset builder (a graph visualization is depicted in Figure 2) using the open-source version of OpenLink’s Virtuoso Triple Store.1 Development of the prototype triple store has been continually informed by our ongoing partnerships with four project teams engaged in separate but related prototyping projects2 to enrich the metadata in the HathiTrust corpus. Through these interactions, the project team encountered the need to accommodate several additional use cases surrounding the selection of materials for worksets and methods for directly enriching bibliographic metadata describing the entities that constitute worksets.

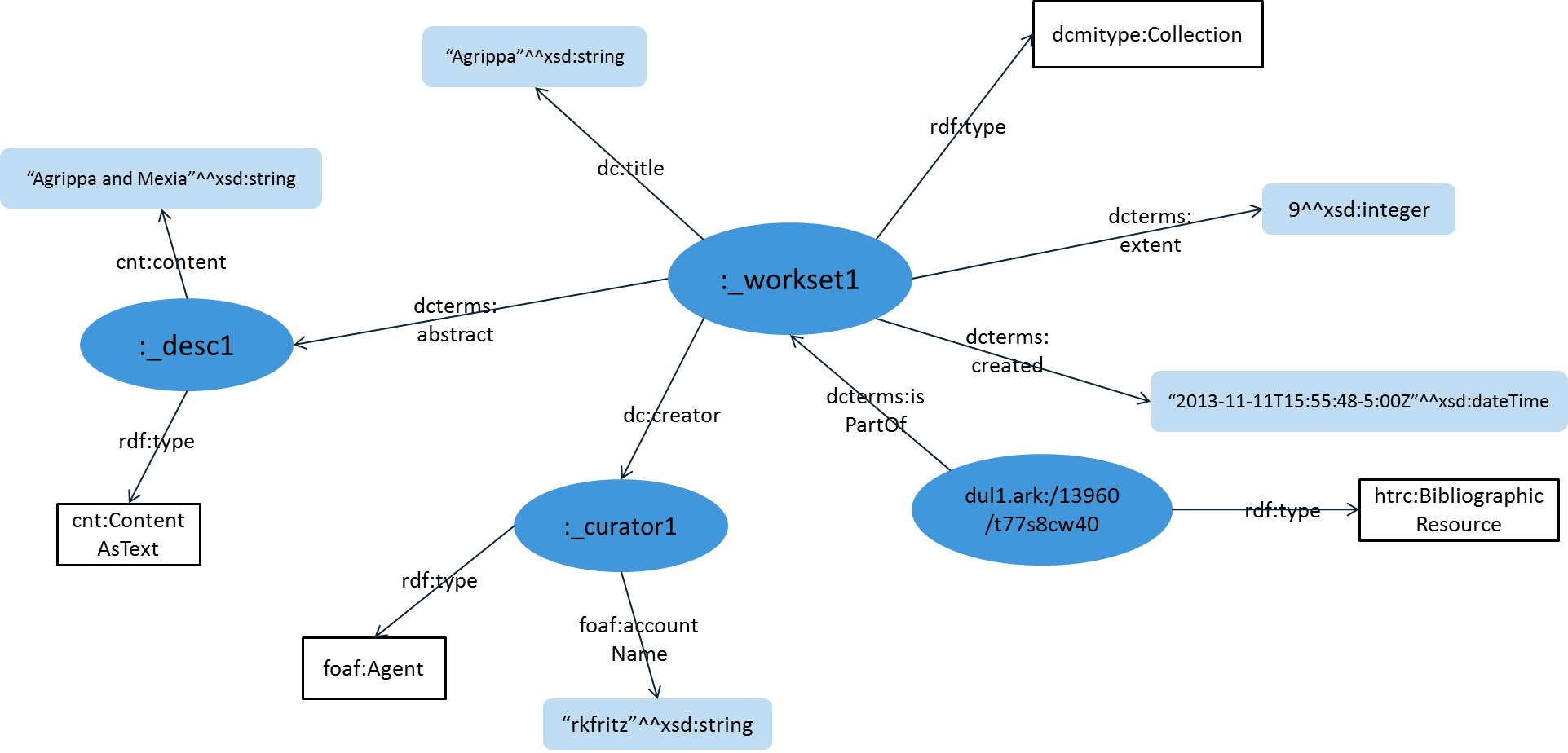


Figure 2. Graph representation of the WCSA Workset model.

In collaboration with the Oxford *ElEPHãT* project, we have explored extensions and adaptations to the bibliographic metadata that describes volumes within the HathiTrust corpus that will facilitate the deduplication process for scholars as they gather research materials, enabling them to remove redundant resources from their worksets more efficiently. We are also working with a team at the Maryland Institute for Technology in the Humanities to explore the best way to leverage annotations of bibliographic metadata. This latter case exploits the RDF-based Open Annotation standard3 as a means for enriching bibliographic metadata without making direct changes to values already recorded within the original MARC metadata records.

**Future Work**

We are currently engaged in exploring additional extensions to the prototype’s underlying data model in order to more fully address the need for more fine-grained units of analysis, as identified by Fenlon et al. (2014). The need to consider page-level rather than volume-level content has already informed the use of new metadata description entities that better characterize pages of digitized content as bibliographic artifacts. Utilizing previous work on arbitrary segmentation of web-based resources (Sanderson, Ciccarese, and Van de Sompel, 2013), we are currently formalizing methods by which finer grained sub-page features—paragraphs, sentences, and other page fragments—can reliably be identified and exploited as workset members in their own right. Complementary methods for identifying and leveraging literary forms such as music and poems, among others, are also under development.

# **Notes**

1. http://virtuoso.openlinksw.com.

2. http://worksets.htrc.illinois.edu/worksets/?p=101.

3. http://www.openannotation.org/spec/core/.

# **References**

**Fenlon, K., Senseney, M., Green, H., Battacharyya, S., Willis, C. and Downie, J. S**. (2014). Scholar-Built Collections: A Study of User Requirements for Research in Large-Scale Digital Libraries. Paper for presentation at the *77th ASIS&T Annual Meeting,* Seattle, WA, 31 October–5 November 2014.

**Henry, C. and Smith, K.** (2010). Ghostlier Demarcations: Large-Scale Text Digitization Projects and Their Utility for Contemporary Humanities Scholarship. In *The Idea of Order: Transforming Research Collections for 21st-Century Scholarship*. Council on Library and Information Resources, pp. 106–15.

**Sanderson, R., Ciccarese, P. and Van de Sompel, H.** (2013). Designing the W3C Open Annotation Data Model. *Proceedings of the 5th Annual ACM Web Science Conference,* Paris, 2–4 May 2013.

**Varvel, V. E. J. and Thomer, A.** (2011). Google Digital Humanities Awards Recipient Interviews Report (CIRSS Report No. HTRC1101). Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, IL.