Conceptualizing worksets for non-consumptive research

Jacob Jett, University of Illinois at Urbana-Champaign
Chris Maden, University of Illinois at Urbana-Champaign
Colleen Fallaw, University of Illinois at Urbana-Champaign
Megan Senseney, University of Illinois at Urbana-Champaign
J. Stephen Downie, University of Illinois at Urbana-Champaign

Abstract
The HathiTrust (HT) digital library comprises 4.5 billion pages (composing 12.9 million volumes). The HathiTrust Research Center (HTRC) – a unique collaboration between University of Illinois and Indiana University – is developing tools to connect scholars to this large and diverse corpus. This poster discusses HTRC's activities surrounding the discovery, formation and optimization of useful analytic subsets of the HT corpus (i.e., workset creation and use). As a part of this development we are prototyping a RDF-based triple-store designed to record and serialize metadata describing worksets and the bibliographic entities that are collected within them. At the heart of this work is the construction of a formal conceptual model that captures sufficient descriptive information about worksets, including provenance, curatorial intent, and other useful metadata, so that digital humanities scholars can more easily select, group, and cite their research data collections based upon HT and external corpora. The prototype's data model is being designed to be extensible and fit well within the Linked Open Data community.
Keywords: Conceptual Models, RDF, Digital Humanities, HathiTrust, Linked Open Data, HathiTrust Research Center
Citation: Jett, J., Maden, C., Fallaw, C., Senseney, M., Downie, S. (2015). Conceptualizing worksets for non-consumptive research. In *iConference 2015 Proceedings*.

Acknowledgements: The authors gratefully acknowledge the Andrew W. Mellon Foundation for generously funding the Workset Creation for Scholarly Analysis Project; the HathiTrust Research Center, the Center for Informatics Research in Science and Scholarship, and the University of Illinois Library and Graduate School of Library and Information Science for institutional support; and WCSA co-PIs Tim Cole and Beth Plale for project leadership.
Research Data: In case you want to publish research data please contact the editor.
Contact: jjett2@illinois.edu, crism@illinois.edu, mfall3@illinois.edu, mfsense2@illinois.edu, jdownie@illinois.edu

1 Introduction

The HathiTrust Digital Library comprises 4.5 billion pages (composing 12.9 million volumes). The HathiTrust Research Center (HTRC) – a unique collaboration between University of Illinois and Indiana University – is developing tools to connect scholars to this large and diverse corpus. Because more than two thirds of the HathiTrust corpus is under copyright, novel methods for indirect (i.e., non-consumptive) analytic access need to be developed by the HTRC. The Workset Creation for Scholarly Analysis: Prototyping Project (WCSA) is a Mellon-funded initiative of the HTRC. This poster examines one of the questions the WCSA initiative seeks to answer: "How can we best formalize the notion of collections and worksets within the HTRC context?"

Within the context of HTRC and WCSA, we describe "worksets" as a type of collection created by scholars for their research that is specialized to the HathiTrust context and intended to facilitate computational analysis. Leveraging existing data drawn from a study of how potential user groups of the HathiTrust Digital Library create and use collections in their research (Fenlon et al., 2014), we use formal methods to develop a simplified entity-relationship model that describes the essential nature of a WCSA workset. From this model we derive a RDF-based ontological framework around which a prototype triple store is implemented.

2 Discussion

The HTRC is developing a series of workbench services to scholars conducting computational research against the HathiTrust corpus, which includes OCR-generated text and scanned page images, the latter of which often include music, illustrations, maps, or other significant features. Scholars will gather materials of interest both from within and beyond the HathiTrust corpus into a workset, which will then be analyzed via one of HTRC's analytics tools, resulting in a number of data products that can be directly leveraged by the scholar. The data products, themselves, can even be candidates for inclusion in

subsequent worksets. Figure 1 (below) provides a simplified view of the expected scholarly workflow within the HTRC.

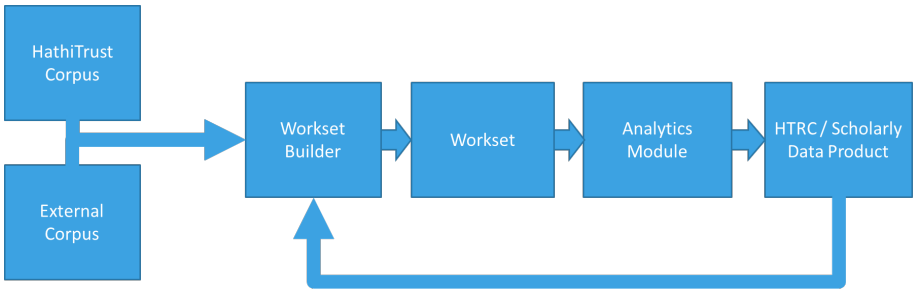

Figure 1: HTRC Scholarly Workflow

A great deal of work on the nature and use of scholarly research collections within the scholarly research cycle already exists (Lynch, 2002; Curral, Moss, & Stuart, 2004; Palmer, 2004; Palmer & Knutson, 2004; Palmer et al, 2006). More recent findings specifically reveal that, while researchers do not necessarily need very large datasets to do interesting work, they do need means to bring heterogeneous objects together into one mass of research materials (Varvel & Thomer, 2011). Scholars have also been found to be aware of the need for architecture that facilitates interoperable use across datasets (Henry & Smith, 2010). Additional studies carried out by WCSA researchers revealed that scholars view their research collections as citable research products in their own right and that they must have sufficient tools to relate their publications back to the sources from which research results are derived (Fenlon et al., 2014).

With regard to text analysis in particular, WCSA research reveals that the ability to segment research materials into highly granular units has direct implications for the kinds of research questions that scholars can explore:

> "Units of analysis are the actual targets of scholars' analytic work: what kinds of things they aim to study, which correspond directly to the kinds of things they aim to collect. … For example, one respondent noted: 'It is very essential to work at the level of a particular chapter, with the actual text…We cannot talk so meaningfully about the work of a writer as a whole, in the abstract. The interpretation is based on actual text, at smaller units of analysis' (P7)." (excerpted from Fenlon et al., 2014).

Previous work on modeling collections within digital spaces also suggests that the relationships between a collection as a distinct entity and the entities that have been gathered into it are quite complex (Renear et al., 2008a and 2008b). Subsequent work demonstrated that semantic web languages like OWL have limitations for expressing these relationships (Wickett, 2009). With both these limitations and the user requirements described above, we developed the following set of functional requirements that the conceptual model and resulting data model must support:

- A workset is a container for a scholar's aggregated units of analysis – analogous to a scholar's research collection;
- A workset is a persistent globally unique entity that can be directly cited;
- A workset possesses provenance properties supporting change awareness within the HTRC context so that a description of its nature at the time of analysis persists over time;
- A workset's membership requirements must be flexible enough to allow for the arbitrary aggregation of heterogeneous resources, with regard to:
    - Granularity of resources that will be considered a unit of analysis and
    - Source from which a particular member entity is retrieved; and
- A workset must contain those properties of its constituent members that propagate to it in support of various types of filtration according to the workset's metadata descriptions, e.g., a workset whose members are all in English is itself in English.

From these functional requirements the conceptual model (Figure 2) was derived.

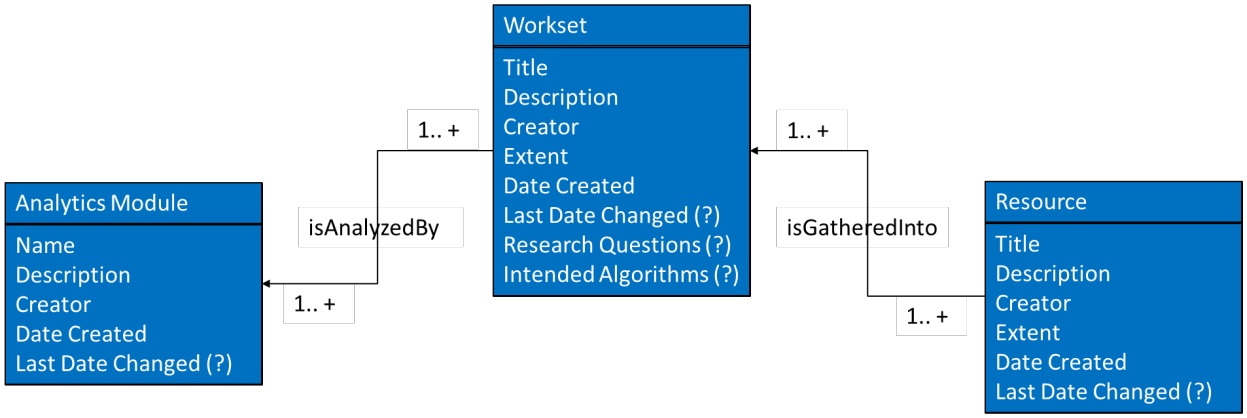

Figure 2: Simple Conceptual Model for Worksets

Using the conceptual model as a basis we have begun developing an OWL schema that captures its nuances. Since a number of useful ontologies have already emerged on the semantic web, we decided to root our initial work in a pre-existing model, specifically the Dublin Core Collections Application Profile.[1] After deciding on an OWL schema that seems to faithfully represent the Dublin Core ontology,[2] we made additional extensions that were critical for our representation. A visualization of the resulting data model can be seen in Figure 3 (below).

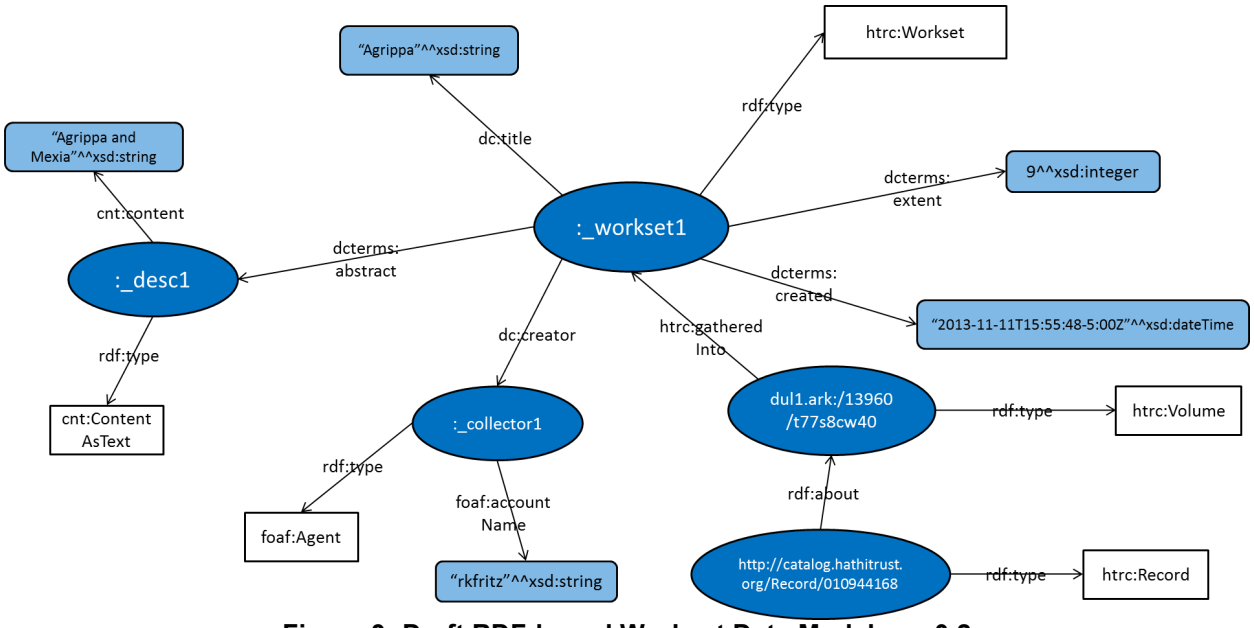

Figure 3: Draft RDF-based Workset Data Model – v. 0.2

The primary features of the data model are the htrc:Workset entity, which is a sub-class of the dcmitype:Collection entity, and the htrc:gatheredInto property, which is a sub-type of the dc:isPartOf property. Other workset specific metadata is captured using the appropriate Dublin Core properties, e.g., dc:creator, dc:created, dc:title, etc.

The data model supports each of the mentioned functional requirements save for supporting arbitrary granularity of a workset's member entities. At the current stage of development, the existing data model only supports entities that correspond to whole volumes or whole pages of volumes. We have implemented the WCSA Workset OWL using the open source version of OpenLink's Virtuoso Triple Store.[3] We are using this prototype test bed to refine the data model's performance and to add additional

[1] http://dublincore.org/groups/collections/collection-ap-summary/2007-03-09/
[2] http://purl.org/NET/dc_owl2dl/dcam
[3] http://virtuoso.openlinksw.com

extensions, such as those that will be needed to support arbitrary segmentation of resources. Regarding this latter issue – segmentation of resources – past web ontology work describes at least one method by which it can be accomplished (Sanderson, Ciccarese, & Van de Sompel, 2013).

3 Conclusion

The conceptual work described in this poster and the resulting prototype being implemented from it demonstrate methods and means by which the HTRC can realize infrastructure that better aligns with scholarly needs. Present work has revealed shortfalls within the HTRC's existing infrastructure. Among other issues, the HathiTrust corpus does not currently support persistent and unique identifiers for page-level entities, and many of the existing analytics modules do not adequately support the range of resources and levels of granularity on which scholars would ideally like to carry out analyses. The preliminary workset model as currently implemented within the HTRC requires further development to fully meet the requirements described above. For example, it is does not currently accommodate heterogeneous source materials and subdivision within those materials.

A prototype triple store is being used to assess the data model's performance, especially with regard to adequate query response times. We are in the process of developing additional entities that support arbitrary segmentation of resources to better meet the needs of our scholarly users. We have developed abstract entities that represent content at the page-level. The HTRC is in the process of realizing these abstractions through the creation of additional infrastructure focused on page-level granularity by minting identifiers for various representations of page-level content. Once page-level considerations have been addressed, we will be exploring methods for identifying finer-grained sub-page features, including paragraphs, sentences, images, and individual words. We are also working on complementary methods for identifying literary forms such as poems, recipes, music, and lyrics, among others.

We are also considering how best to model the more complex property value propagation relationships between the workset entities and their member entities. While no best method has yet been advanced to accomplish this task, there are clear benefits to developing a means as the information can be leveraged to provide additional contextualizing information about the scholar's workset, which in turn can increase the ease with which peer review processes might proceed. We will continue refining the model and exploiting it to upgrade the prototype's functionality and performance over the coming months so that the HTRC can leverage lessons learned and new infrastructure developments.

References

Currall, J., Moss, M., & Stuart, S. (2004). What is a collection? *Archivaria 58*, 131-146.

Fenlon, K., Senseney, M., Green, H., Battacharyya, S., Willis, C., & Downie, J. S. (2014). Scholar-built collections: A study of user requirements for research in large-scale digital libraries. Paper presented at *The 77th ASIS&T Annual Meeting*. (Seattle, WA, Oct. 31 – Nov. 5, 2014).

Henry, C., & 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* (pp. 106‑115). Council on Library and Information Resources.

Lynch, C. (2002). Digital collections, digital libraries, and the digitization of cultural heritage information. *First Monday, 7*(5).

Palmer, C. L. (2004). Thematic research collections. In Schreibman, S., Siemens, R., and Unsworth, J. (Eds.) *A Companion to Digital Humanities*. Blackwell Publishing, Oxford.

Palmer, C. L., & Knutson, E. (2004). Metadata practices and implications for federated collections. *Proceedings of the 67th ASIS&T Annual Meeting* (Providence, RI, Nov. 12-17, 2004).

Palmer, C. L., Knutson, E., Twidale, M., and Zavalina, O. (2006). Collection definition in federated digital resource development. *Proceedings of the 69th ASIS&T Annual Meeting* (Austin, TX, Nov. 3-8, 2006).

Renear, A. H., Wickett, K. M., Urban, R. J., and Dubin, D. (2008a). The return of the trivial: Formalizing collection/item metadata relationships. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries 2008* (Pittsburgh, PA, June 16-20, 2008).

Renear, A. H., Wickett, K. M., Urban, R. J., Dubin, D., and Shreeves, S. (2008b). Collection/Item metadata relationships. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008* (Berlin, Germany, Sept. 22-26, 2008).

Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2013). Designing the W3C open annotation data model. *Proceedings of the 5th Annual ACM Web Science Conference* (Paris, France, May 2-4, 2013).

Varvel, V. E. J. & 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.

Wickett, K. M. (2009). Logical expressiveness of semantic web languages for bibliographic information modeling. *Proceedings of the 2011 iConference* (Seattle, WA, Feb. 8-11, 2011).