Workset Creation for Scholarly Analysis
Preliminary Research at the HathiTrust Research Center

J. Stephen Downie[1], Tim Cole[2], Beth Plale[3], John Unsworth[4]
[1]Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign; [2]University Libraries, University of Illinois at Urbana-Champaign;
[3]School of Informatics and Computing, Indiana University; [4]Brandeis University

Introduction

Scholars rely on library collections to support their scholarship. Out of these collections, scholars select, organize, and refine the worksets that will answer to their particular research objectives. The requirements for those worksets are becoming increasingly sophisticated and complex, both as humanities scholarship has become more interdisciplinary and as it has become more digital.

The HathiTrust's computational infrastructure is being built to support large-scale manipulation and preservation of these representations, but it organizes them according to catalog records that were created to enable users to find books in a building. *These catalog records were never meant to support the granularity of sorting and selection or works that scholars now expect, much less page-level or chapter-level sorting and selection out of a corpus of billions of pages.*

HathiTrust Research Center

The HathiTrust Research Center (HTRC) is the research branch of the HathiTrust (HT). The HathiTrust aims to be the official digital archive of the world's most important libraries. There are more than eighty partners in HathiTrust, and membership is open to institutions throughout the world.

See **http://www.hathitrust.org/data**.

DESCRIPTION	Count
Total Volumes	10,644,397
Public Domain Volumes	3,305,946
Book Titles	5,598,627
Serial Titles	277,216
Pages	3,725,538,950
Disk Memory in Terabytes	477
Linear shelf distance in miles	126
Original material weight in tons	8,649

Table 1: The magnitude of the ever-growing HathiTrust corpus.

MARC FIELD	Percent of Sample Records
245 Title Statement	> 99%
260 Publication Distribution, etc.	92%
500 General Note	41%
650 Topical Term / 653 Index Term – Uncontrolled	39% / 13%
050 LC Classification No / 082 Dewey Classification No	17% / 13%
655 Index Term -- Genre Form	12%

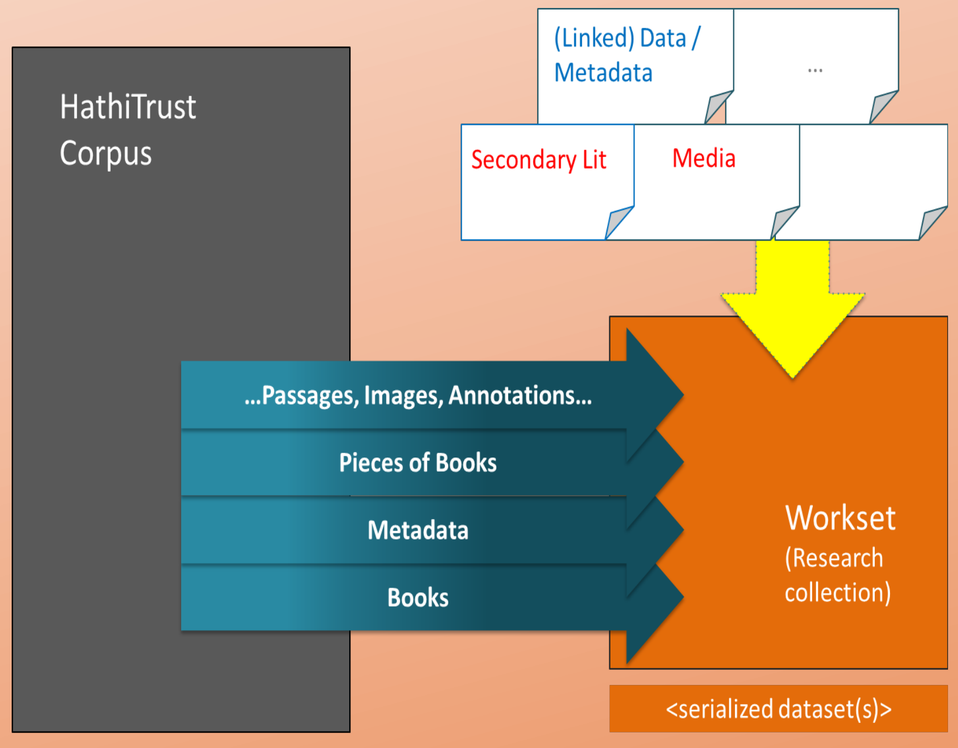
MARC FIELD	Percent of Sample Records
650 Topical Term	6%
050 LC Classification No / 082 Dewey Classification No	27% / 4%
655 Index Term -- Genre Form	5%

Tables 2a and 2b: MARC field population across two sampled collections.

Principal Research Questions

The ***Workset Creation for Scholarly Analysis: Prototyping Project*** (WCSA) seeks to address three sets of tightly intertwined research questions regarding:

1. Enriching the metadata describing the HathiTrust corpus through mining of the resources themselves and leveraging end-user annotations;
2. Augmenting string-based metadata with URIs to leverage external services and Linked Open Data to facilitate discovery and the process of organizing HathiTrust resources into collections and worksets; and,
3. Formalizing the notion of collections and worksets in the context of the HathiTrust Research Center.



Workset Questions

1. How can we best formalize the notion of collections and worksets within, and beyond , the HTRC context?
2. What are the necessary elements of a "collection"? What are the necessary elements of a "workset"?
3. How can we balance rigor with extensibility and flexibility?
4. What roles do "data", "metadata", "annotations", "tags", "feature sets", and so on, all play in the conception, creation, use and reuse of collections and worksets?

Key Outcomes

1. A set of prototype algorithms that could be used by scholars to define new collections for analysis.
2. A collection of new metadata outputs from algorithms that could be used to assist in improving access to the HathiTrust corpus and/or be used in novel analyses within or beyond the HTRC.
3. A suite of prototype Linked Open Data (or similar) resources that will expose the outputs of the prototype algorithms for use both within and beyond the HTRC context.
4. A formal model of collections and worksets that can be used to shape the development of new discovery, search and analytic resources both within and beyond the HTRC context.
5. The realization of the formal collection model in a form that can be used to encode collections and worksets for use in actual HTRC analyses
6. Recommendations based upon the experience of creating the first five outcomes listed above designed to guide both the HTRC and the digital scholarship community in formulating a high-impact, long-term research and development plan.

What is a Workset?

1. A workset is an aggregation of materials brought together for the purpose of analysis.
2. Worksets are conceptual and must be expressible in a variety of ways
   - Need to allow creation outside of HathiTrust
   - Need facilitate inclusion of resources beyond HathiTrust
   - Need to facilitate the inclusion of resources at many different levels of granularity beyond the book
3. Worksets encapsulate the specific materials that underwent analysis.
   - Need to capture provenance information
   - Possible recording of parameters
4. Worksets should be able to spawn descendants but otherwise immutable

Some Motivating Questions

Worksets could contain (inspired by participants of 2012 HTRC UnCamp):

1. Volumes pertaining to Japan / in Japanese
2. All volumes relevant to the study of Francis Bacon
3. Music scores or notation extracted from HT volumes
4. Images of Victorian England extracted from HT vols.
5. Volumes in HT similar to TCP-ECCO novels
6. 19th century English-language novels by female authors
7. Representative sample (by pub date & genre) of French language items in HT

Project Workstreams

1. **Workset Structures and Formal Semantics**
   - Work housed at the Center for Informatics Research in Science and Scholarship at the Graduate School of Library and Information Science
2. **Workset Prototyping Projects**
   - Four projects funded by the grant but conducted by community teams
   - Build proof-of-concept tool to help answer one or more motivating questions from community
   - $40,000 for each team for roughly one year
   - Request for Proposals (RFP) to be released in November 2013
   - Finalist meeting in Chicago, mid-February 2014
   - Project awards announced, mid-March 2014
   - Projects demonstrated, Spring 2015

Acknowledgments

Thanks to M. Senseney, K. Fenlon, C. Willis, K. Wickett, P. Organisciak, H. Green and S. Bhattacharyya. Special thanks to project funder, The Andrew W. Mellon Foundation.

The Andrew W. Mellon Foundation