

HathiTrust Research Center

Workset Creation for Scholarly Analysis (WCSA)

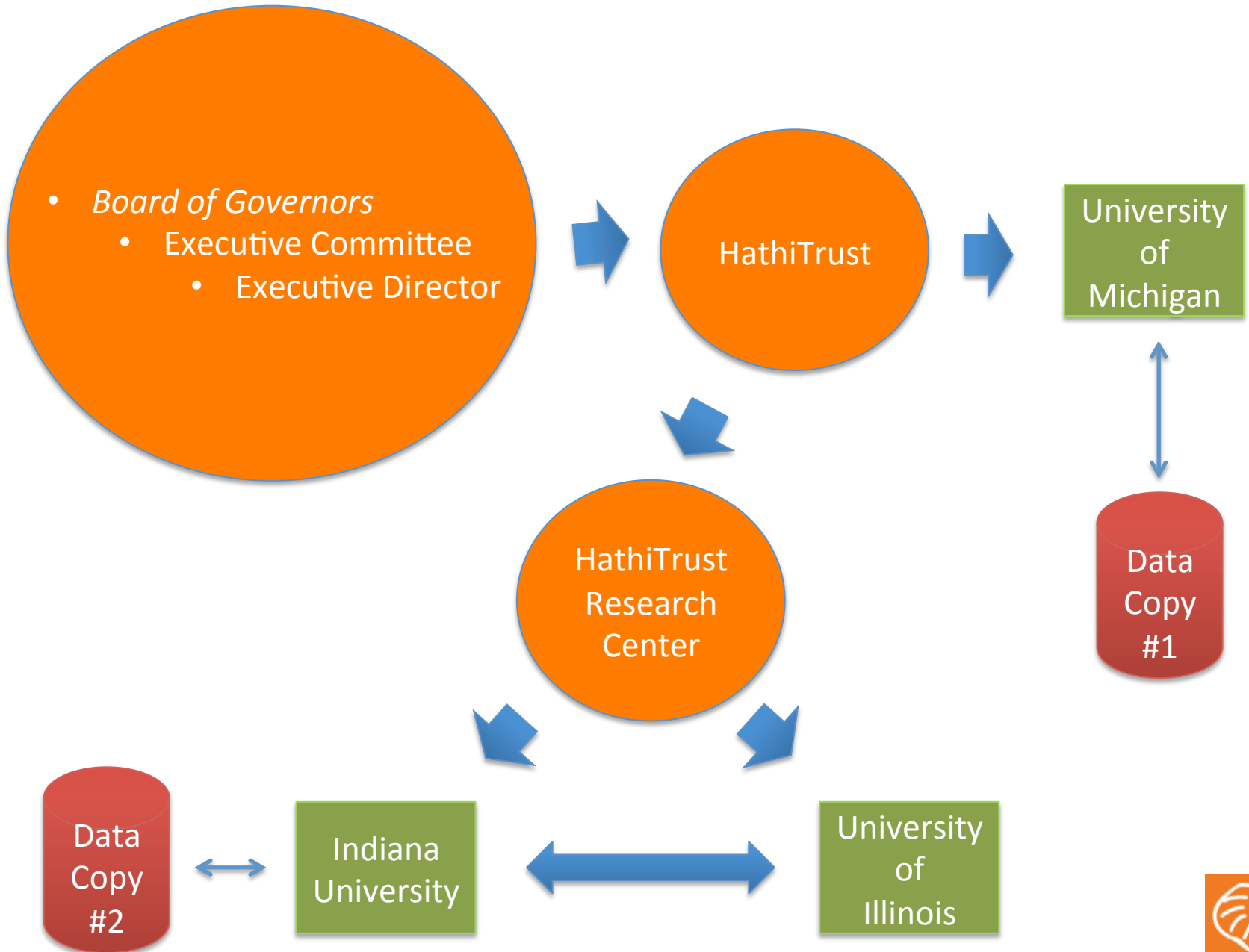


<http://bit.ly/worksets>

Workset Creation for Scholarly Analysis: Prototyping Project

- Collection analysis and prototype tools & services to facilitate work-set creation
 - J. Stephen Downie, Tim Cole, Beth Plale
 - Andrew W. Mellon Foundation
 - 1 July 2013 - 30 June 2015
- Proposal Narrative:
 - <http://bit.ly/htrcwcsa>





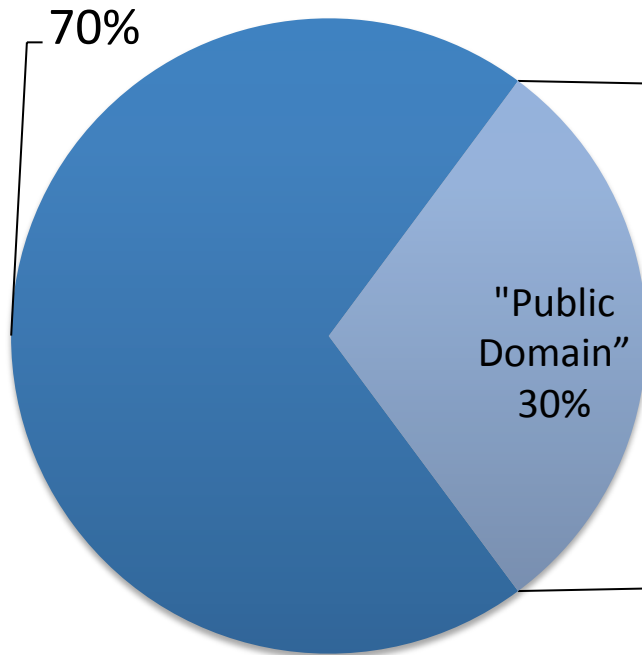
HathiTrust “Wow” Numbers

- 10,800,458 total volumes
- 5,658,812 book titles
- 281,890 serial titles
- 3,780,160,300 pages
- 484 terabytes
- 128 miles
- 8,775 tons
- 3,454,586 volumes (~32% of total) in the public domain



Content Distribution

In-copyright or
undetermined

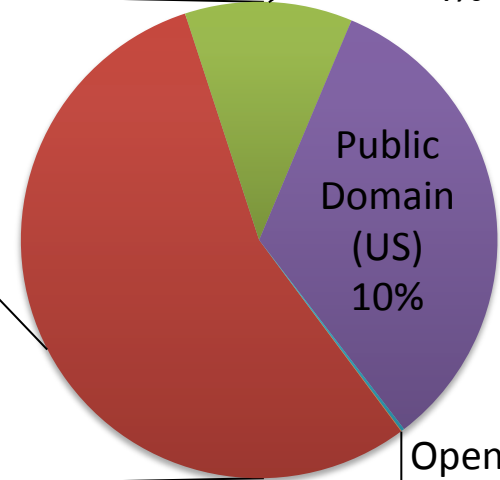


U.S. Federal
Government
Documents
(worldwide)

4%

Public Domain
(worldwide)

15%



Creative Commons

.01%



Grand Motivation

The ability to slice through a massive corpus constructed from many different library collections, and out of that to construct the precise workset required for a particular scholarly investigation, is an example of the “game changing” potential of the HathiTrust...



Motivation & Models

Collections, corpora, work sets,:

- Aggregations of items brought together in some context:
 - Archival
 - Curatorial
 - Experimental
 - Referential
 - Thematic (for research)



Carl Spitzweg. 1850
The Bookworm (Der Bücherwurm)

Analogy: HathiTrust workset for analysis as the contents of a scholar's carrel in a library



Dimensions of Workset Creation (Illustrative)

My work-set should contain (inspired by 2012 UnCamp):

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



HathiTrust Collection Builder

out Collections Help Feedback Hi Megan Finn Senseney! My Collections

HATHI TRUST Digital Library

FULL-TEXT CATALOG

100

255

Private Public

Cancel Add

Collection can be searched

Find a collection

1-50 of 1468 of all collections

Previous 1 2 3 4 ... 30 Next


2 items last updated: 10/11/10

erations'

r: quoddy

Featured Collection

Records of the American Colonies



Published documents--leg court proceedings, record: correspondence, etc.--from original colonies and their predecessors.



MARC Metadata Shortcomings I

MARC Field	Percent of records in OCLC having instance of this field
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%

Table 1. Frequency of MARC fields in OCLC Records



MARC Metadata Shortcomings II

MARC Field	Percent of British Novel MARC records having instance of this field
650 Topical Term	6%
050 LC Classification No / 082 Dewey Classification No	27% / 4%
655 Index Term -- Genre Form	5%

Table 2. Frequency of MARC fields used in 2,386 descriptions of 19th century British novels digitized from UIUC collections



Why Worksets?

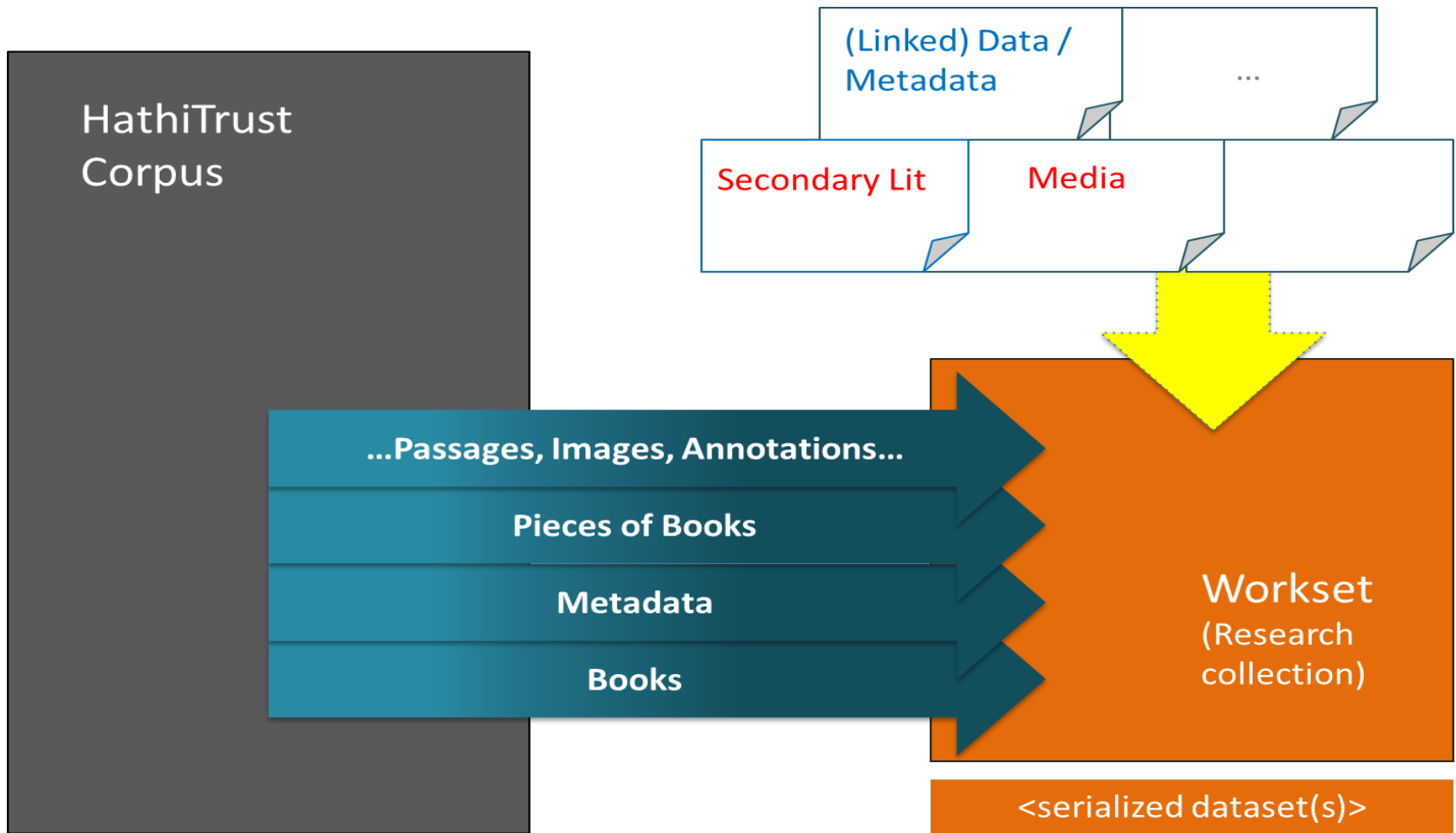
- The result of a first-level, rough filter
- Better scale for intensive analytics
- Provides essential scope for certain analytics
 - (e.g., word frequency scope over Bacon's essays)
- Some tools (are trained to) work best on a narrow, homogeneous work-set
- Eliminate noise that would otherwise arise by asking questions across whole of HT



What is a Workset?

- A workset is an aggregation of materials brought together for the purpose of analysis.
 - Beyond HathiTrust
 - Beyond volumes
- Worksets are conceptual and need to be expressible in a variety of ways
 - Need to allow creation outside of HathiTrust
- A workset encapsulates the specific materials that underwent analysis.
 - Provenance
 - Possible recording of parameters
 - Worksets should not/cannot be retrospectively modified





Conceptualizing a Workset



Research Questions (Illustrative only)

- Can we enrich the HathiTrust corpus metadata by distilling analytics over full text?
- Can we augment string-based metadata with URIs for recognized entities – e.g., names, subjects, publication location, etc. -- and by doing so can we leverage external services to facilitate discovery and clustering of resources?
- Can we leverage existing, well-defined external corpora to identify complementary subsets of HT volumes, and having done so can we demonstrate the ability to create and perform analytics over an integrated workset that includes resources external to HT?



Key Workset Questions

- Can we formalize the notion of collections and worksets in the HTRC context?
- What are the necessary elements of a “collection”?
What are the necessary elements of a “workset”?
- How can we balance rigor with extensibility and flexibility?
- 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?



Two Project Streams

- Workset formal structures and semantics
 - Work in conjunction with Center for Informatics Research in Science and Scholarship at the Graduate School of Library and Information Science
- WCSA Prototyping Projects
 - Four projects funded by the grant but conducted by community teams



WCSA Timeline

- July 2013: Project Start
- Q1: User needs assessments / focus groups
- Q2: HT Corpus characterization
Request For Prototype Proposals
- Q3: RFP Finalist Meeting, February 20, Chicago, IL
Prototype experiment funding awarded
- Q4-6: Prototype experiments done
Metadata workflow & work-set modeling
- Q7-8: Planning for prototype to production
Report out
- June 2015: Project ends



WCSA Activities to Date

- Conduct User-needs assessments
 - Digital Humanities in Lincoln, Nebraska
 - ACM/IEEE Joint Conference on Digital Libraries in Indianapolis, Indiana
 - HTRC UnCamp 2013
- Evaluate metadata content for HathiTrust Corpus and build representative testbed
- Release RFP: <http://bit.ly/wcsarfp>



Prototype Grants

As part of project, HTRC will make 4 sub-awards

- \$40K awarded to each of 4 non-HTRC teams
- HTRC will collaborate with each team
 - Access to representative test data / metadata set
 - Collaborate on work with HT / HTRC APIs, etc.

RFP & Sub-Award Schedule	
2013-11-22	RFP Available (http://bit.ly/wcsarfp)
2013-12-16	Letters of Intent Due (preferred)
2014-01-13	Final Proposals Due
2014-02-20	Finalist Meeting
By 2014-03-15	Award Notification for projects running April-Dec, 2014



Importance of Mellon IP Agreement

- Absolutely essential that subprojects be able to sign and fulfill the Mellon IP agreement.
- Nothing can proceed unless agreed and signed.
- Basically no room for negotiation.



Conclusion

- Worksets are fundamental to the scholarly computational analysis enterprise
- We need a better understanding of their:
 - Constituent parts
 - Creation
 - Manipulation
 - Use and reuse
- Prototypes to lead to deeper tool development and metadata enhancement



Questions?

