RESEARCH CENTER

Using Collections and Worksets in Large-Scale Corpora: Preliminary findings from the Workset Creation for Scholarly Analysis prototyping project (WCSA)

Harriett E. Green, UIUC ; Katrina Fenlon, UIUC ; Megan Senseney, UIUC ; Sayan Bhattacharyya, UIUC ; Craig Willis, UIUC ; Peter Organisciak, UIUC ; J. Stephen Downie, UIUC ; Timothy Cole, UIUC ; Beth Plale, IU

Research question

How do researchers, especially humanities scholars, use collections in the course of their research, particularly in the context of textual corpora?

Motivation

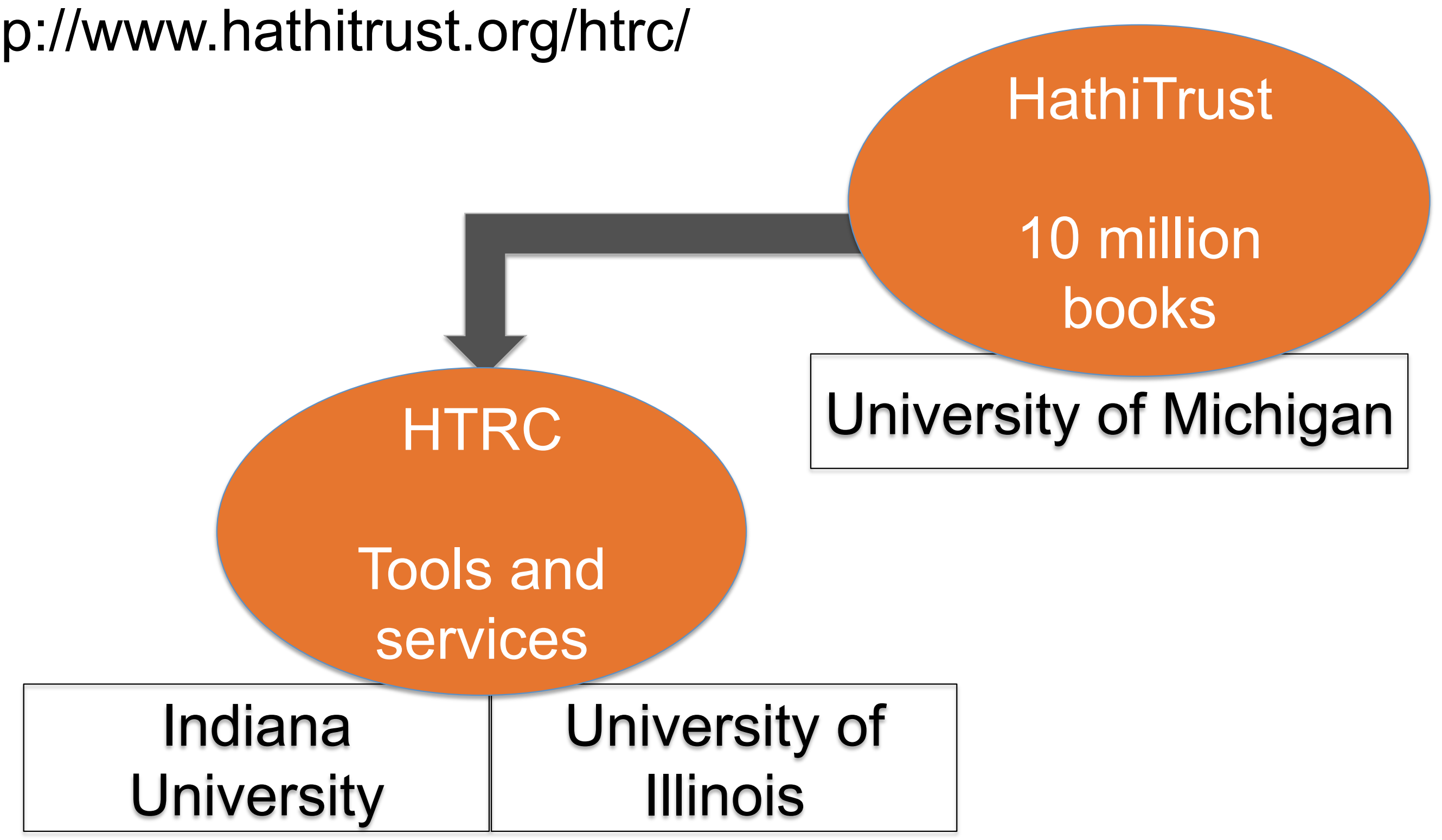
- Scholars increasingly use digitized primary sources
- Scholars rely on different kinds of collections in the course of their research
- Worksets enable scholars to develop collections of digitized sources for computational analysis within and beyond the HathiTrust corpus

Methods

- Semi-structured focus groups and interviews
- 13 focus group participants ; 5 individual interviews
- Humanities scholars, digital humanities researchers, and librarians and service providers
- Conducted at Digital Humanities 2013, Joint Conference on Digital Libraries 2013, & HathiTrust Research Center UnCamp 2013
- Multiple rounds of qualitative coding using AtlasTI 7 for inter-coder reliability

The HathiTrust Research Center

The HathiTrust is a repository of over 10 million volumes (3 billion pages) of text. The HathiTrust Research Center (HTRC) is the research branch of the HathiTrust. The HTRC offers a suite of tools and services, which enable computational access to the HathiTrust corpus. http://www.hathitrust.org/htrc/



Preliminary findings

"…we need ways to slice this book. So we need to slice it by page…by poem… We potentially need to slice it by sections within a poem…"

"they use a lot of corpus configurations… Subcorpus building…And partitions-building…So this is for contrastive analysis"

"We have words, text units, and intermediate structure. Those three levels hold different types of properties"

"Books are often not interesting without knowledge of the logical works or units within…"

Theme 1: Roles of collections

- Collecting and workset-building are basic scholarly activities, but are often unacknowledged labor

"collection-building is scholarly activity… we also need to think about how to document…the labor that goes into and the kinds of knowledge that go into the decisions in making a collection, and the knowledge that's gained from that process."

"the valorization of corpus-building…The recognition at the scientific level"

"[If] I have a corpus and nobody is allowed to see it but wonderful things come out of it… That's not really research… We are tying to get accountability for the kind of work we are doing. And it's important for us to show the basis of our work."

Theme 2: Units of analysis

- Objects of analysis must be highly divisible, pieces must be identifiable, movable, and connectable to highly granular metadata

"Collaborative curation… You could create the data collaboratively, and then explore them collaboratively"

"The book is not a unit of great interest – you want all the poems that aren't listed in the metadata. The metadata from the library is very coarse, especially in respect to the goal you have. There's no opportunity for the experts to provide the deep metadata to share in the broad infrastructure that librarians do very well."

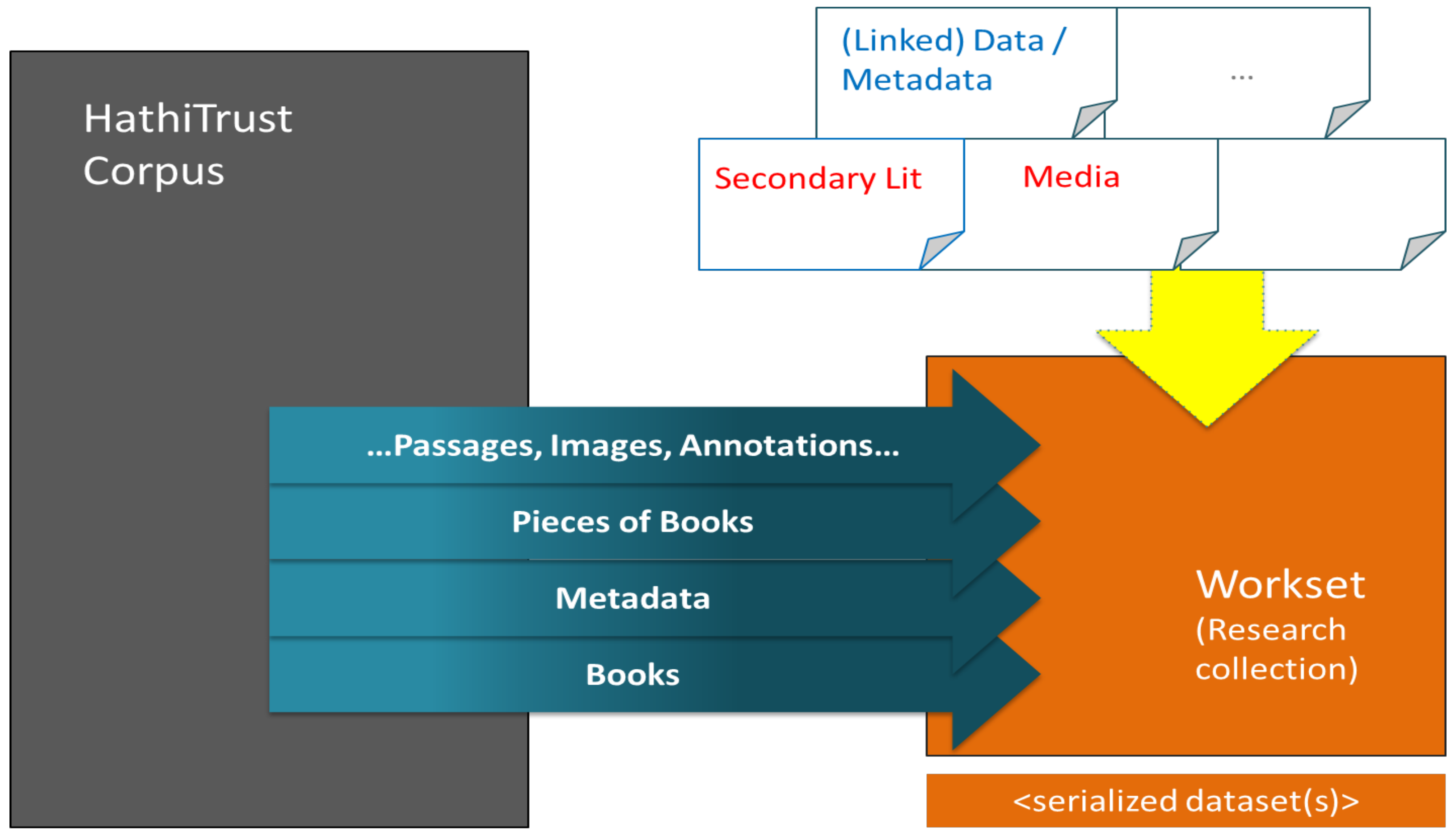
"you've done all this work, and you then have the authoritative metadata. You have the best metadata in the world, and no one will take that from you. Because it has not been blessed."

Theme 3: Better metadata

- Scholars want non-bibliographic, expert-enriched, shareable metadata

What is a workset?

Machine-actionable, referential research collection in the HathiTrust context.



WCSA objectives

- Enable routine computational analysis across subsets of materials in the HathiTrust corpus
- Engage scholars in tool design
- Enrich metadata in the HathiTrust corpus
- Formalize the notion of worksets

Next steps

- Formalize workset model to allow researchers to identify, select, and pull together subsets of texts within massive corpora
- Implement worksets in HathiTrust Research Center
- Employ preliminary findings about user requirements to inform further tool-building prototyping projects to be awarded by WCSA

Acknowledgements

CIRSS
GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
UNIVERSITY LIBRARY
The Andrew W. Mellon Foundation