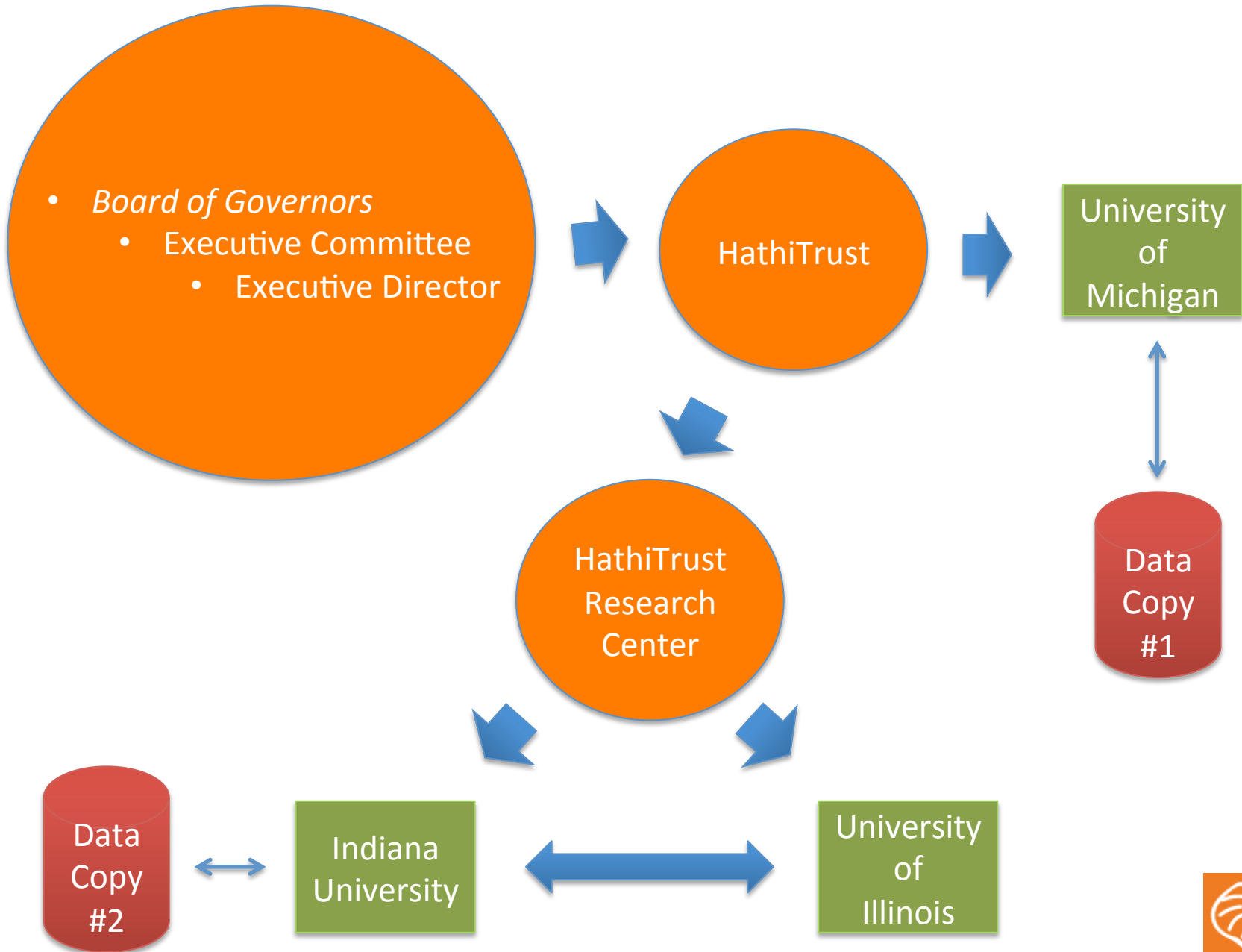


HathiTrust Research Center

Workset Creation for Scholarly Analysis (WCSA)





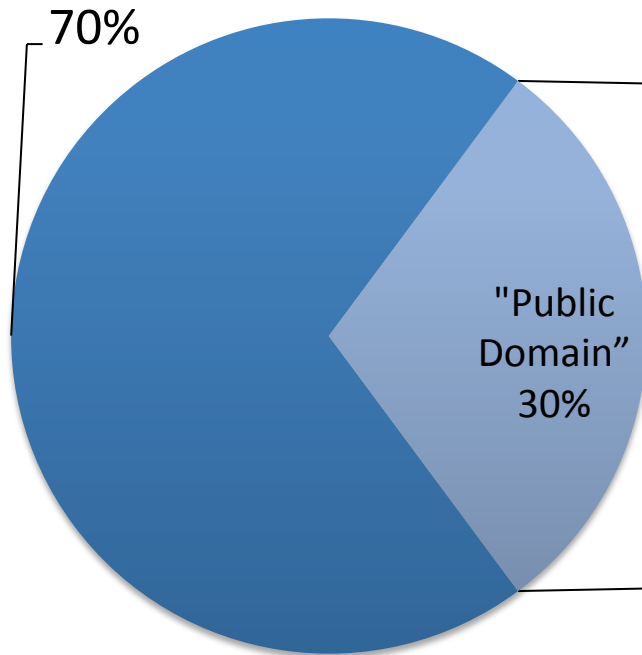
HathiTrust “Wow” Numbers

- 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

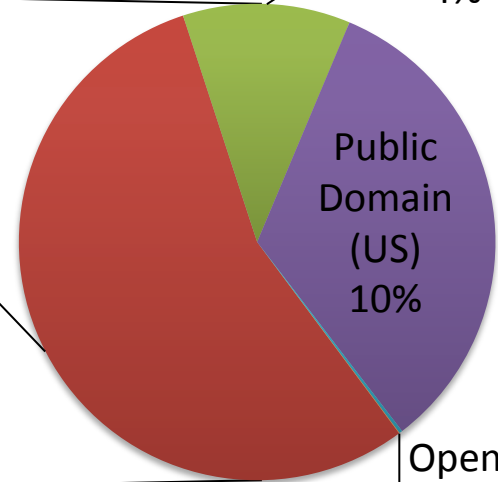


U.S. Federal
Government
Documents
(worldwide)

4%

Public Domain
(worldwide)

15%

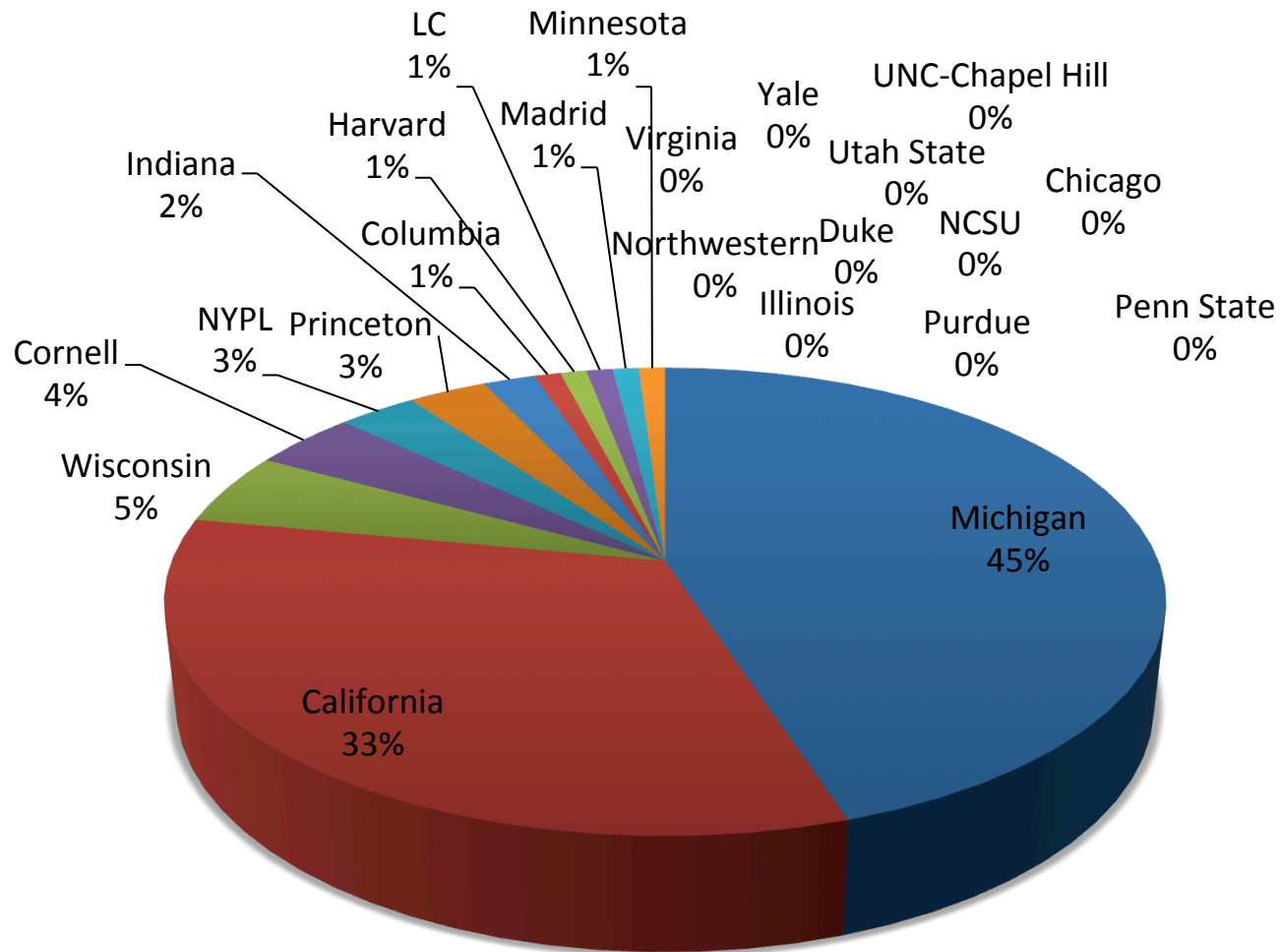


Creative Commons

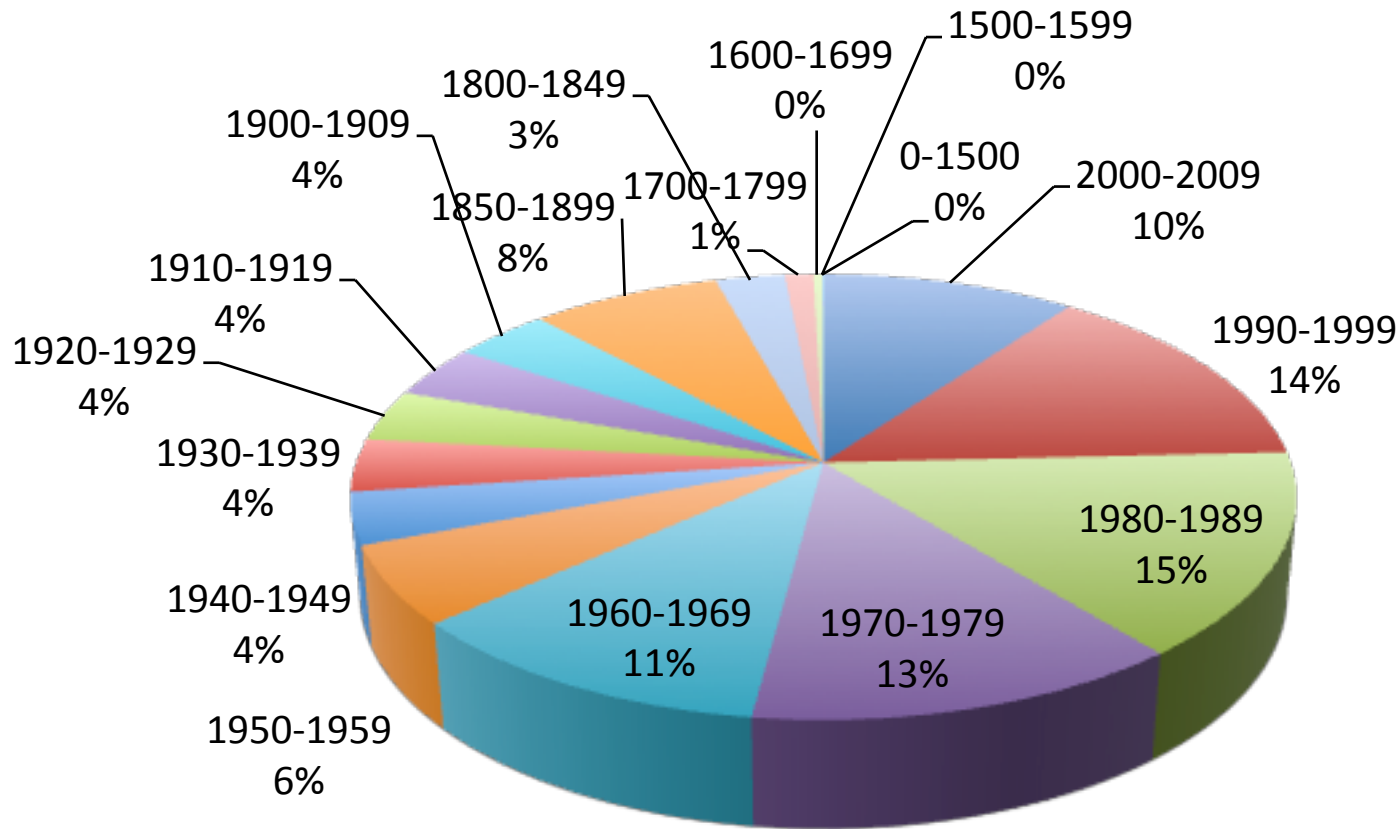
.01%



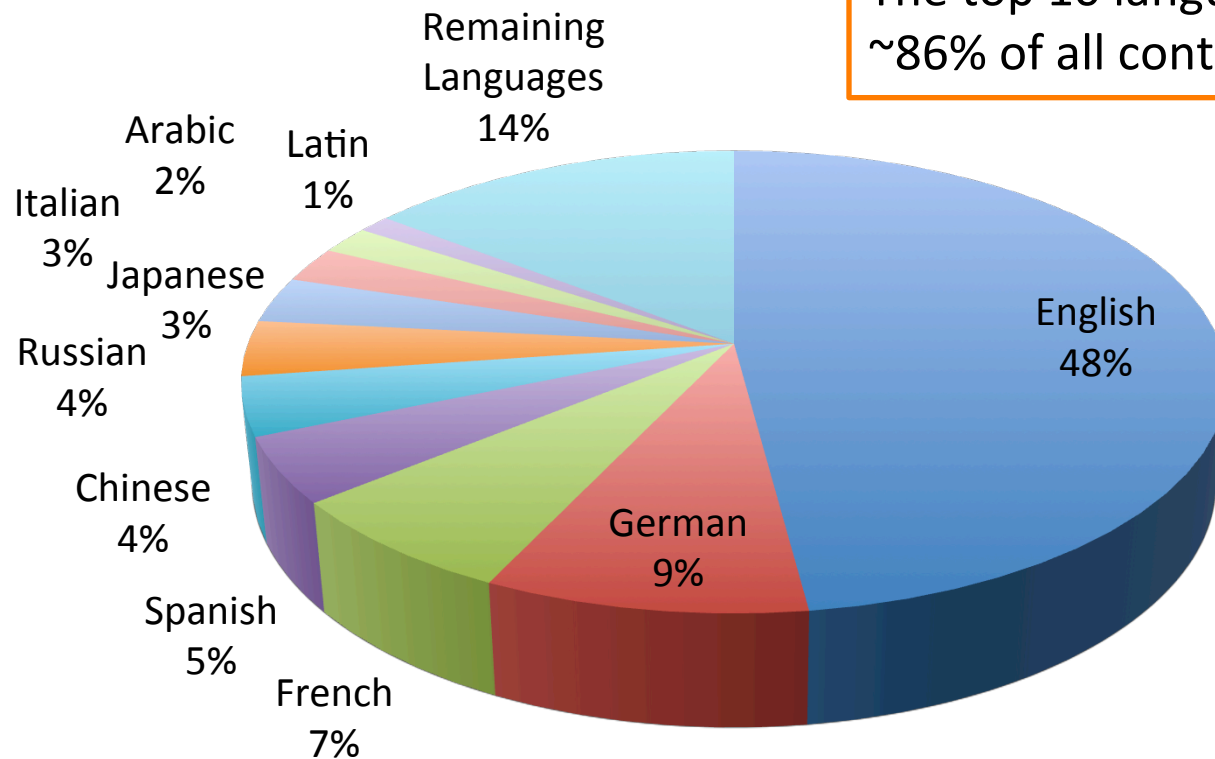
Content Sources



Dates



Language Distribution



The top 10 languages make up ~86% of all content



Research 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...



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



This isn't enough...



HathiTrust Collection Builder

out Collections Help Feedback Hi Megan Finn Senseney! My Collections

HATHI TRUST Digital Library

FULL-TEXT CATALOG

Search [] Full view only

Collection Name: Your collection name (100)

Description: Add your collection description. (255)

Private Public

Cancel Add

Collection can be searched

[Create new collection](#)

Find a collection []

Recently Updated

ons with at least (all items)

Collection Title []

g 1-50 of 1468 of all collections

Previous 1 2 3 4 ... 30 Next


erations'

r: quoddy

2 items
last updated: 10/11/10

Featured Collection

[Records of the American Colonies](#)



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



This isn't enough either...



Workset Creation for Scholarly Analysis: Prototyping Project

- Collection analysis and prototype tools & services to facilitate work-set creation
 - Principal investigators:
 - J. Stephen Downie, Tim Cole, Beth Plale
 - Andrew W. Mellon Foundation
 - 1 July 2013 - 30 June 2015



WCSA Timeline

- July 2013: Project Start
- Q1: User needs assessments / focus groups
- Q2: HT Corpus characterization
Request For Prototype Proposals
- Q3: RFP Finalist Workshop (Chicago) February 20?
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



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



Why Worksets?

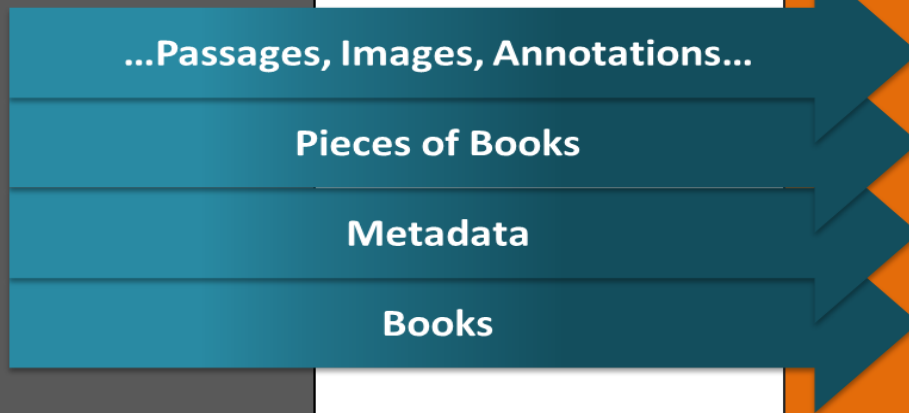
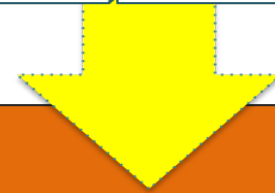
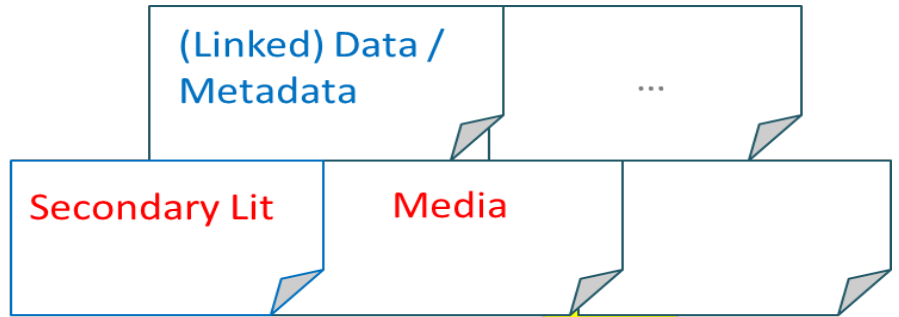
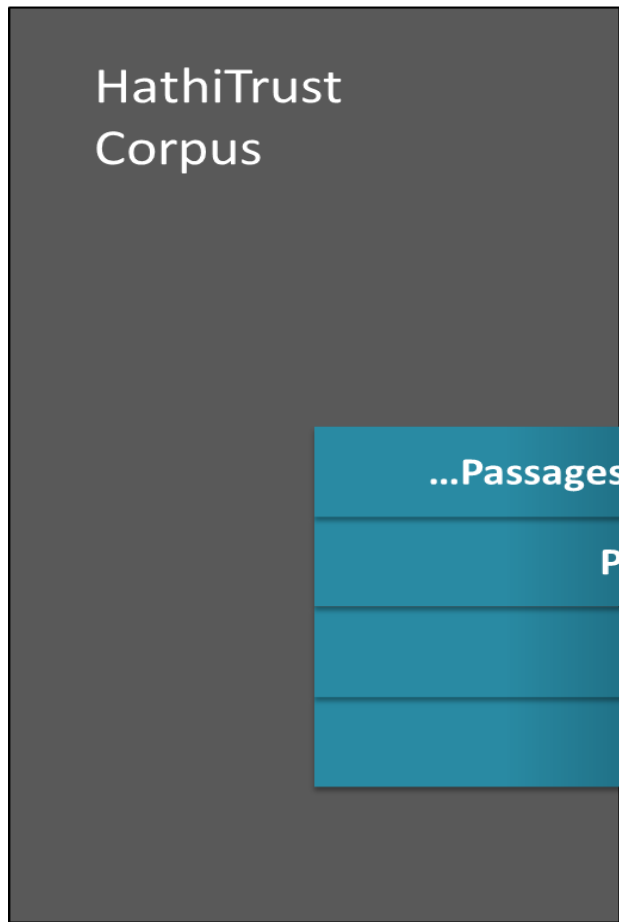
- The result of a first-level, rough filter
- Better scale for intensive analytics
- Provides essential scope for certain analytics
 - Word frequency scope over Bacon's essays
- Some tools (are trained to) work best on a narrow, homogeneous work-set
- Eliminate noise that would otherwise arise by asking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





Scope



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?



Content Evaluation

- Understand corpus coverage
- Anticipate metadata enrichment
- Construct representative sample



Who Are Our Researchers?

- Humanities scholars? Computer programmers and Technologists? Digital humanities research teams?
- Previous research in scholarly use of digital resources (Duff and Cherry 2000; Brockman et al. 2001; Warwick et al., 2008; Sukovic, 2008 and 2011; RIN 2011)
- Identify use cases for HTRC and large-scale, digitized text corpora



GOOGLE DIGITAL HUMANITIES AWARDS RECIPIENT INTERVIEWS REPORT

- 2011 report prepared for the HTRC by UIUC researchers at GSLIS – Varvel and Thomer
- Interviewed researchers awarded Google Digital Humanities Research Awards
- Findings for scholarly requirements/needs included improved metadata and accurate OCR scans
- Available to download at <http://www.hathitrust.org/htrc>



Feedback from UnCamp 2013

My work-set should contain...

- 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





Scholarly Requirements

We are interested in understanding how scholars and researchers that use digital book and serials collections decide which texts (or parts of texts) to include in collections used for analysis. This includes:

- How researchers identify, select and obtain access to texts to include in their analysis
- Understanding the specific fields/disciplines that work with these sources along with the types of research questions and analysis applied.
- Desired units of analysis (works, manifestations, pages, n-grams OCR, images, etc.)
- Transformation and preprocessing steps;
- Understanding sources and criteria used for identifying texts
- Specific methods of selection
- Methods of analysis
- Challenges to working with these digital collections (e.g., OCR quality, duplication)



Focus Groups and Interviews

- Goal: To understand practices of humanities researchers using digital collections, especially in the context of large-scale text corpora
- Participants:
 - Humanists
 - Researchers working with humanists
 - Information professionals and librarians
- Timeline: Summer 2013-February 2014



Study Design

1. General types of data, materials, or collections
2. Purposes of collections
3. Selection or inclusion/exclusion criteria
4. Sources, acquisition, and access
5. Pre-processing and analysis
6. Post-analysis
7. Challenges



Methodology: Qualitative content analysis of user responses



Goal: To identify common themes and patterns in users' responses

- An approach based on inductive reasoning to condense raw data (responses from users in interviews and focus groups) into categories and themes
- “Directed” approach:
 - Coding starts with a “theory” of themes/ categories that are expected to emerge from the qualitative data
 - Themes/categories refined during coding process



Coding manual (currently under development)

- Coding manual consisting of category names, rules for assigning codes, and examples
- Data: drawn from HTRC Uncamp 2012, HTRC Uncamp 2013, JCDL 2013 focus groups and interviews
 - Transcriptions of audiorecordings (ongoing)
 - Survey instrument queried users about their experiential practices of organizing datasets
 - Fairly structured questions



Selected categories from coding manual

- Challenges — access rights
- Challenges — OCR quality
- Collections — comprehensiveness
- Objects — data
- Objects — genre
- Sources — Google Books
- Sources — Selection Criteria — Language
etc.



Selected examples for categories

- Category:
Challenges— Access Rights
 - User: “I check to see if a volume has substantial copyrighted text included in it already as quotes or extracts”
 - Category: Objects — Temporal
 - User: “Classic materials”
 - User: “single-authored books of poetry between 1840 and 1900”
- Etc.



Participant Demographics

- Positions:
 - Junior and senior faculty at liberal arts colleges and universities
 - Computer programmers
 - Librarians
 - Data scientists
 - Academic technologists
 - Graduate students
- Domains:
 - English literature, classics, linguistics, library and information science, and history
- Institutions:
 - Academic institutions in Great Britain, Singapore, Germany, France, and United States



Early Findings

- Roles of collections
- Need to implement granular, actionable units of analysis
- Importance of expert-enriched, shareable metadata



“collection-building is scholarly activity... we also need to think about how to document not just the status of different versions but also 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.”

“Today it is viewed as something very technical to prepare a corpus. But I think it’s getting more and more... interesting to do. And one day, it will be unrelated to technical stuff, and it will get closer to something of value.”

“the valorization of corpus-building...The recognition at the scientific level”

“I’m learning a lot through this organizing of my material and it’s informing what will be the main argument of my research”

“[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 trying to get accountability for the kind of work we are doing. And it’s important for us to show the basis of our work.”

Figure 1. Selected focus group and interview excerpts on collection- and workset-building.



“...we need ways to slice this book. So we need to slice it by page...We need to slice it by poem, which doesn't conveniently overlap or match the page boundaries. We potentially need to slice it by sections within a poem...”

“they use a lot of corpus configurations, like subcorpora. Subcorpus building... And partitions-building. Partition is to slice the corpus in parts, the sum of which is the whole. So this is for contrastive analysis”

“Books are often not interesting without knowledge of the **logical works or units within...”**

“that's a whole different dicing intellectually ... Being able to support the huge variety of those kinds of ways of thinking about [texts] at that logical level is a bit challenging. But I think it's one that somehow has to be approached...”

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

Figure 2. Selected focus group and interview excerpts on divisibility and objects of analysis.



“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.”

“Collaborative curation... You could create the data collaboratively, and then explore them collaboratively”

“one thing is getting the data out. But then the next step is, 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.”

“it would be very important to have the ability to say [of the metadata], this is wrong ...having a workflow which supports that would be important. So the whole idea of **social addition** comes really into play here.

Figure 3. Selected focus group and interview excerpts on metadata enrichment and sharing.



How does this fit in with WCSA?

Current Phase of Project:

- Analysis from focus groups and interviews will...
 - Help set priorities for technical development and creation of pilot services for scholars
 - Inform revisions to RFP and evaluation of responses
- Content evaluation will...
 - Determine aspects of current metadata that require improvement
 - Identify corpus strengths, which will further inform outreach efforts
 - Inform creation of representative 100k-volume test bed to be used for testing prototyping projects

Next Phase of Project:

- WCSA Prototyping Projects
 - Four projects funded by the grant but conducted by community teams
- 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



Situating the HTRC workset idea within the overall landscape of the (Digital) Humanities

Claim: The challenges posed by the idea of the HTRC workset are similar to the challenges of the (digital) humanities



Digital humanities as “integrative”

- “Humanities computing” (a narrowly focused, tool-oriented approach) became, as it evolved into digital humanities, an “integrative endeavor.” (Zorich 2008).

Zorich, Diane. “A survey of digital humanities centers in the United States.” CLIR Publication No. 143, Council on Library and Information Resources, 2008.



What ~~are~~ were the humanities (like)? (In the era of modernism/modernity?)

- The humanities (at least, in their *modernist* version) ~~are~~ were “integrative”:
 - Knowledge production in the humanities happens through
 - *generating insights* by **integrating** *disparate (oppositional)* material together; and
 - *creating meaning* through such *integration*.
 - Two paradigmatic examples (from the modernist canon):
 - Walter Benjamin on historical insight
 - Sergei Eisenstein on montage



Benjamin on historical insight

- “The true picture of the past flits by. *The past can be seized only as an image which flashes up at the instant when it can be recognized...*”
- “To articulate the past historically does not mean to recognize it ‘the way it really was’ (Ranke). It means to seize hold of a memory as it flashes up at a [present] moment.”

(Walter Benjamin, *On the Concept of History*, 1937)



Eisenstein on montage

- “In nature we never see anything isolated, but everything in connection with something else which is before it, beside it, under it, and over it.” — Goethe
- “ ‘Montage’ is an idea that arises from the *collision* of independent shots, even *opposite* to one another: the dramatic principle.”

(Eisenstein in his essay “Film Form”, 1949)



What are the humanities like, today? (In postmodernism/postmodernity?)

The **postmodern condition**:

- Disappearance of binary oppositions
(e.g. between “inside” and “outside”)
- “Incredulity towards metanarrative”
- Smoothness, convergence, frictionless-ness

These are all also *integrative*, but in a different way.

(Through regularities rather than irregularities?)



The “modernist” thematic collection (“digital study carrel”) versus a “postmodernist” HathiTrust Workset?

- The “digital study carrel”/contextually massive thematic collection
 - Large datasets sliced and recombined together, in highly contingent ways, as happens in the scholar’s “study carrel” (Martin Mueller’s metaphor)
 - A heterogeneous site of encounters of opposites
 - A *modernist* space?
- The HathiTrust workspace
 - A homogeneous site for mining with algorithmic regularity
 - A *postmodernist* space?



The (slightly old) thematic research collection: a “modernist” space?

- “*Monuments and Dust*, a thematic collection focused on Victorian London... The premise is that *the aggregation of diverse sources* – images, texts, numerical data, maps, and models – will *seed intellectual interaction* by making it possible to discover *new* visual, textual, and statistical relationships within the collection and between lines of research.” (Palmer 2004)
- “Thematic collections ‘concerned with the construction of knowledge from sources of different types, scattered across different subject areas’” (Fraser 2000)



The (new) workset concept (a substrate for text mining): a postmodernist space?

- The emphasis shifts to regularity, coherence, smooth compositionality, rather than contingent encounters between heterogeneities?
 - “Service composition framework which allows users to pull together resources easily in a work environment” (Dempsey 2006)
 - “Coherent aggregation of heterogeneous but thematically associated content” (WCSA proposal)



Users torn between desire for personalized, contingent, “modernist” construction and automated-analogical construction of worksets?

- User comments from interviews/focus groups:
 - “How do I gather works *similar to those I currently have in hand?* Can I *define different kinds of similarity?*”
 - “How do I merge a *HathiTrust collection of works and metadata* with *my set of works and tags* and *my colleague’s annotations?*”



Discussion Questions

- Key questions to look for in the data
- Alternative approaches and methodologies
- Knowing what we know about user needs to date, what are the implications for formalize the notion of workset
- How does this translate across domains? (e.g., Workset-like objects in science and elsewhere...)
- What are the re-usability and re-productibility implications for such highly individualized and complex digital objects



Thank you!

