HathiTrust Research Center: Improving Scholarly Inquiry

Timothy W. Cole (*t-cole3@illinois.edu*)
Harriett Green (*green19@illinois.edu*)

With slides and other contributions from Stephen Downie, Beth Plale, Colleen Fallaw, Megan Senseney, Katrina Fenlon, et al.

CNI Fall 2013 Membership Meeting
Washington, D.C.
9 December 2013

RESEARCH CENTER

Outline

- The HathiTrust Digital Library (HT)
- The HathiTrust Research Center (HTRC)
- The Workset Creation for Scholarly Analysis (WCSA) Project
- User needs & requirements
- Characterization of bibliographic metadata for corpus
- More about the WCSA RFP & prototyping projects

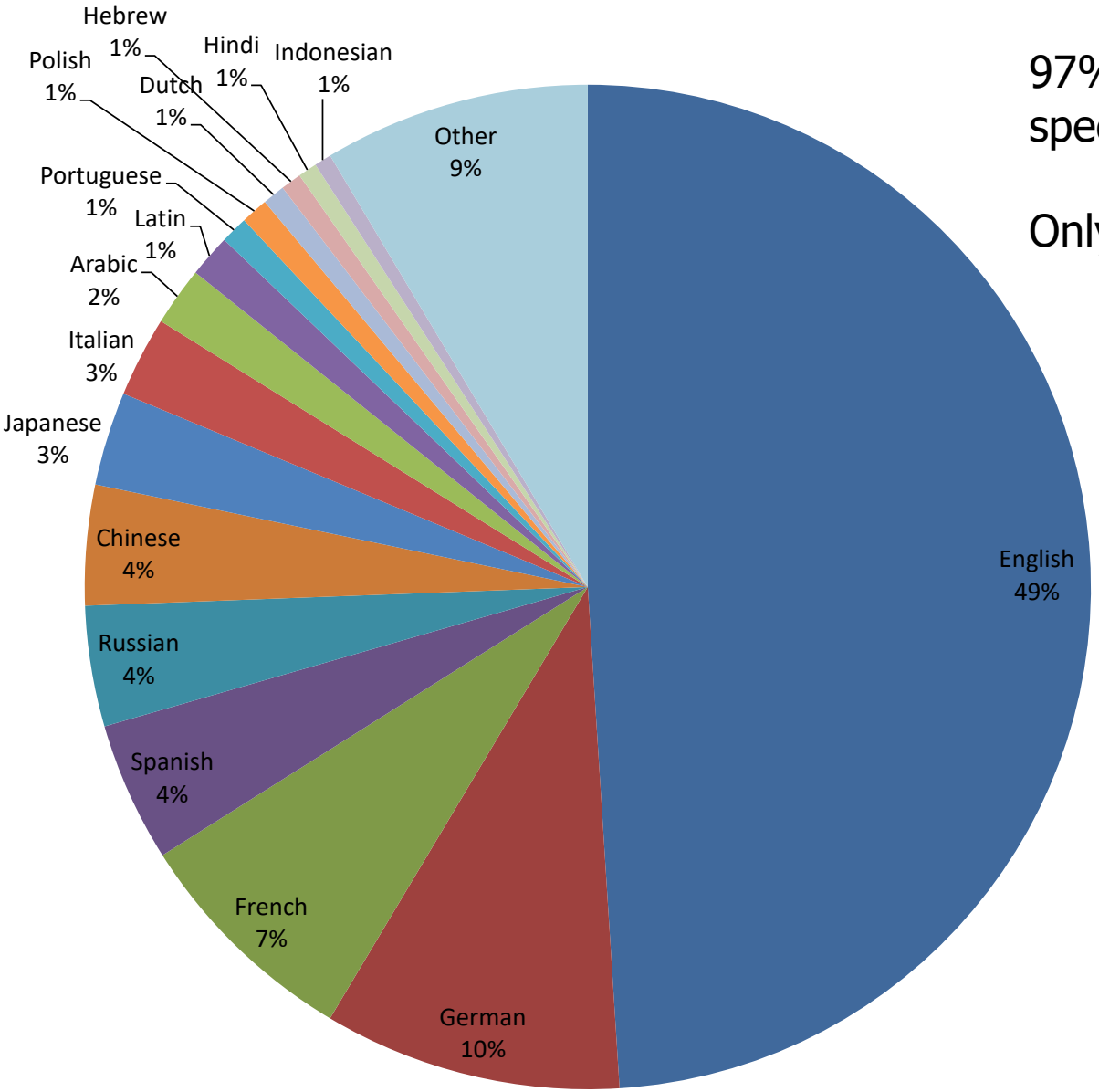
The HathiTrust Digital Library *(hathitrust.org)*

- A digital preservation repository coupled with a highly functional access platform
- An international partnership of 80+ research libraries & consortia
- Provides long-term preservation of and access to volumes of member library collections that have been digitized by Google, the Internet Archive, Microsoft & member institutions
- Currently supports ingest of digitized book and journal content, and similar book-like materials

HT DL by the numbers (as of Nov 2013)

- 10,973,063 total volumes
- 6,067,835 distinct bibliographic items:
  - 5,778,450 book (monographic) titles
  - 289,385 serial titles
- 3,803,630,600 pages
- 487 terabytes
- 3,512,404 volumes (~32% of total) digitized from public domain originals

More than just US Libraries



97% of bibliographic records specify resource language

Only 7% specify more than 1

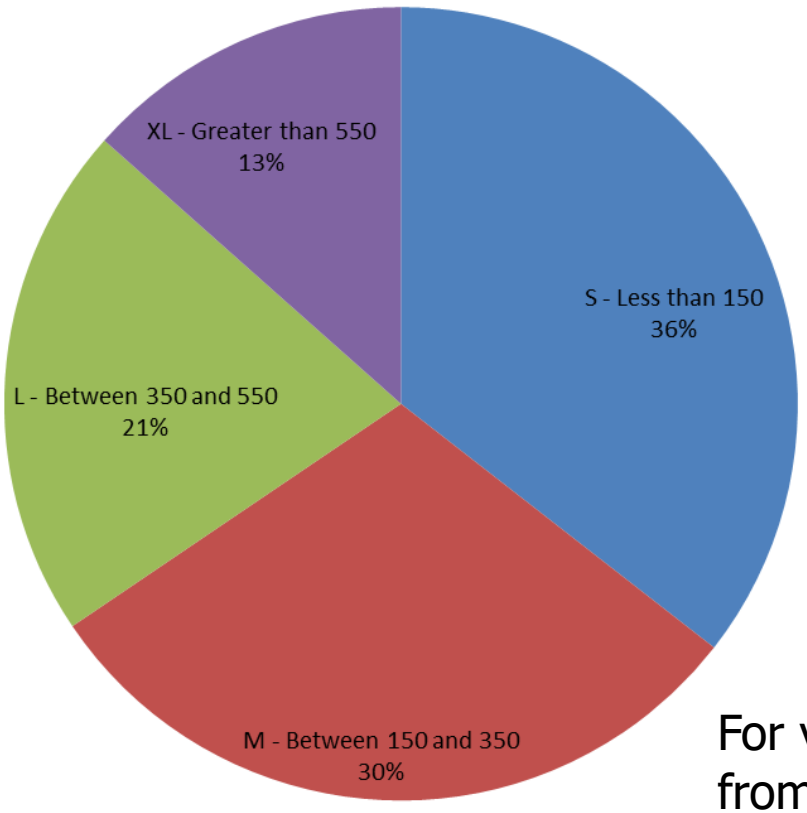
HT DL Searching & Data Availability

- Web User Interface (http://www.hathitrust.org/home):
  - Full text keyword (includes indexed metadata)
  - Bibliographic metadata keyword
  - Advanced (field-specific) bibliographic metadata searching

- Bibliographic metadata (http://www.hathitrust.org/data):
  - OAI-PMH & custom bib API – http://www.hathitrust.org/bib_api
  - HathiFiles (tab delimited metadata) – http://www.hathitrust.org/hathifiles

- Full-text (http://www.hathitrust.org/datasets):
  - ~300,000 digitized volumes in the public domain – contact for bulk download or use API for volume-by-volume access
  - ~ 3,500,000 volumes digitized by Google from public domain – available by arrangement, typically using rsync. Must agree to conditions of use.

How many pages per volume?



For volumes digitized from public domain sources

The HathiTrust Research Center (1)

HTRC is a collaboration between HT, Indiana University and the University of Illinois at Urbana-Champaign

- Goal is to provide **computational** access to researchers: initially to all content digitized from public domain eventually to the entire HT DL corpus
- Currently hosts
  - complete copy of HT metadata
  - copy of OCR of all HT volumes digitized from public domain
  - copy of OCR of all public domain volumes in HT

Supported by the Sloan Foundation, the Mellon Foundation, IU, & UIUC

The HathiTrust Research Center (2)

HTRC end-user access (so far)

- HTRC Portal
  https://htrc2.pti.indiana.edu/HTRC-UI-Portal2/
  Must login; pull-down login menu (upper right) to sign up (free)
- HTRC Workset Builder
  https://htrc2.pti.indiana.edu/blacklight
  Must login to this interface also; same credentials used.
- HTRC Sandbox (contact us)
  http://sandbox.htrc.illinois.edu:8080
  Clone of Portal, but accessing 250,000 digital public domain volumes
  Supports use data api – http://wiki.htrc.illinois.edu/display/COM
  As well as HTRC Solr Proxy api – http://chinkapin.pti.indiana.edu:9994

More: http://www.hathitrust.org/htrc/faq, http://wiki.htrc.illinois.edu/display/OUT

HTRC Portal (as it is now)

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Login

RESEARCH CENTER

Welcome to the HathiTrust Research Center!

The HathiTrust Research Center (HTRC) provides research access to the public domain text of the HathiTrust Digital Library. The HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Library, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face by developing cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge.

The HTRC provides an infrastructure to search, collect, analyze, and visualize the full text of nearly 3 million public domain works and is intended for nonprofit and educational researchers.

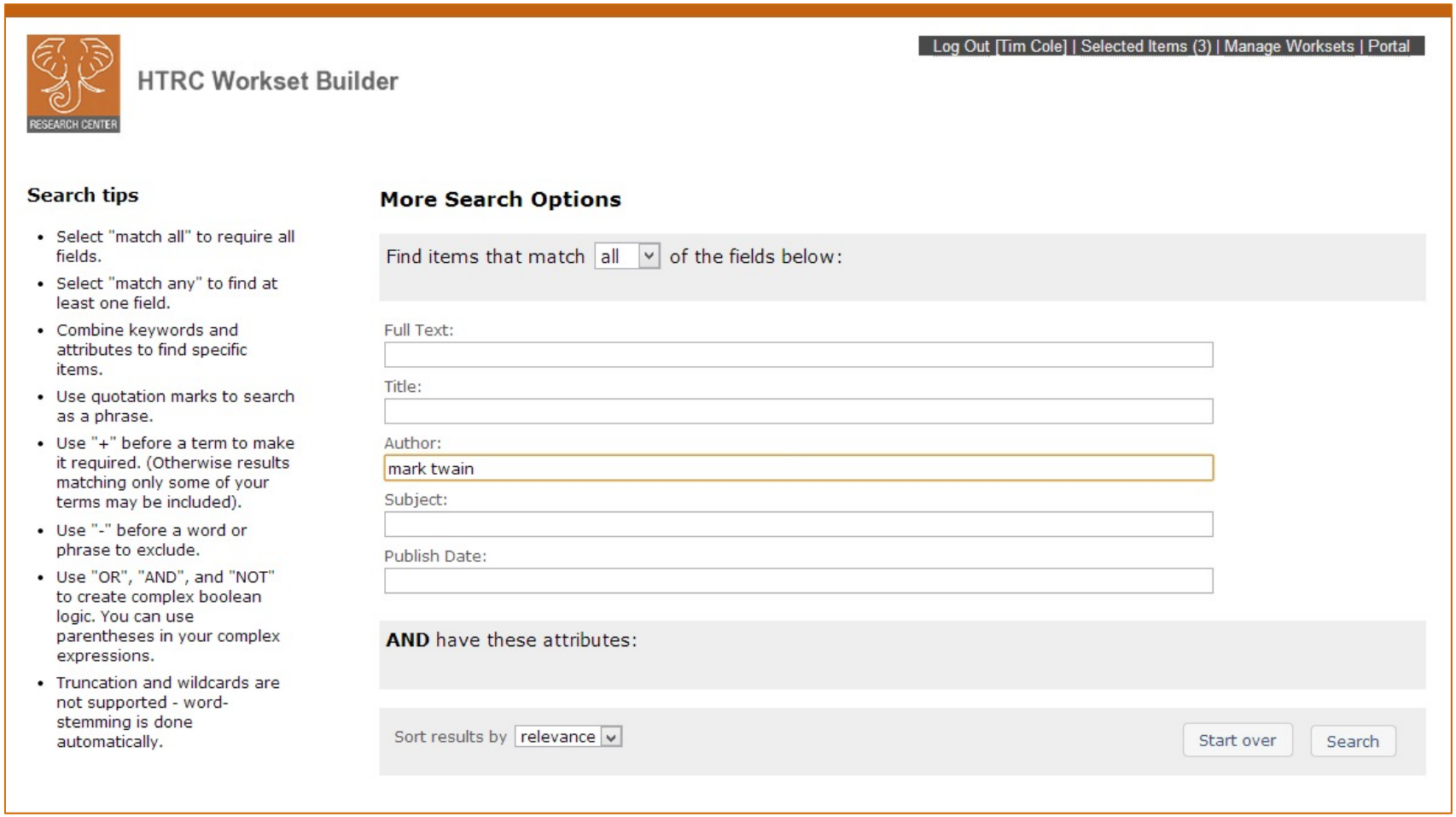
Sign In to Begin

What Can You Do With HTRC Portal?

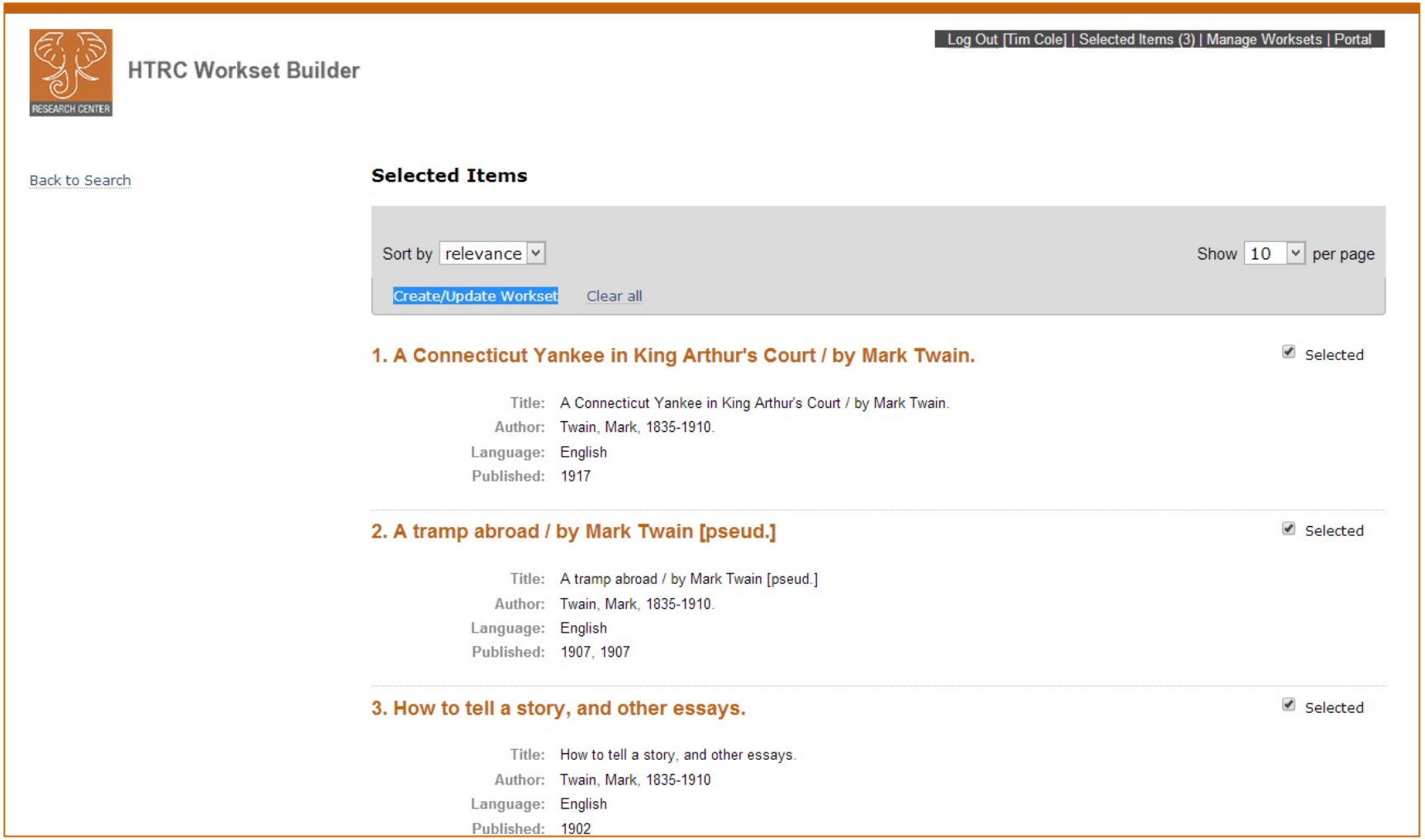
- Create Workset
- Upload Workset
- Browse Workset
- Execute Algorithms

About | Contact

HTRC Workset Builder (as it is now)



Create a small workset / collection



Submit a small workset / collection for analysis

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Signed-in as: tcole3

Available Algorithms

- Marc_Downloader
- Meandre_Classification_NaiveBayes
- Meandre_Dunning_LogLikelihood_to_Tagclou
- Meandre_OpenNLP_Date_Entities_To_Simile
- Meandre_OpenNLP_Entities_List
- Meandre_Spellcheck_Report_Per_Volume
- Meandre_Tagcloud
- Meandre_Tagcloud_with_Cleaning
- Meandre_Topic_Modeling
- Simple_Deployable_Word_Count

Algorithm Parameters

Name: Meandre_Tagcloud

Description: This analysis performs token counts and displays the most frequent tokens in a tag cloud. Counts the tokens for all volumes and displays the top 200 tokens in a tag cloud. No cleaning of the text is performed. NOTE: The volume limit is 1000.

Version: 1.1

Author: Loretta Auvil

Please Input Job Name:

(required)

tc3Test1

Please select a collection for analysis

TwainSmallSample@tcole3

Submit

Your completed and pending analytical jobs

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Signed-in as: tcole3

Active Jobs

Cancel

Job Title	Last Updated	Status	Cancel?
tc3Test2	2013-12-07 13:27:38	Staging	☐

Completed Jobs

Delete Selected | Save Selected

Job Title	Last Updated	Status	Delete/Save?	Saved?
tc3Test1	2013-12-07 13:23:43	Finished	☐	unsaved

About | Contact

Results

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Signed-in as: tcole3

Job Details

Job Title: tc3Test2
Algorithm Name: Meandre_Tagcloud_with_Cleaning
Last Updated: 2013-12-07 13:29:02
Results:

- stderr.txt
- stdout.txt
- tagcloudcleantokencounts.html

Job Parameters:

Name	Value
n_top_tokens	200
stopwords_list_english_url	http://repository.seasr.org/Datasets/Text/common_words.txt
input_collection	TwainSmallSample@tcole3
replacement_rules_url	http://repository.seasr.org/Datasets/Text/ngram_corrections.txt

Job Id:
917614ba-be12-4d2a-8c55-9e625e94c84e
Status:
Finished

View Results

Workset Creation for Scholarly Analysis
Premise

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, worksets, ...:

- Scholars & librarians aggregate items in a variety of contexts:
  - Archival
  - Curatorial
  - Experimental
  - Referential
  - Thematic

  These worksets facilitate, sometimes enable certain kinds of scholarly inquiry

Carl Spitzweg. 1850
The Bookworm
(*Der Bücherwurm*)

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

With apologies to Martin Mueller, et al. 2010. *Towards a digital carrel: A report about corpus query tools*

Anecdotal feedback from UnCamp 2013

My workset 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

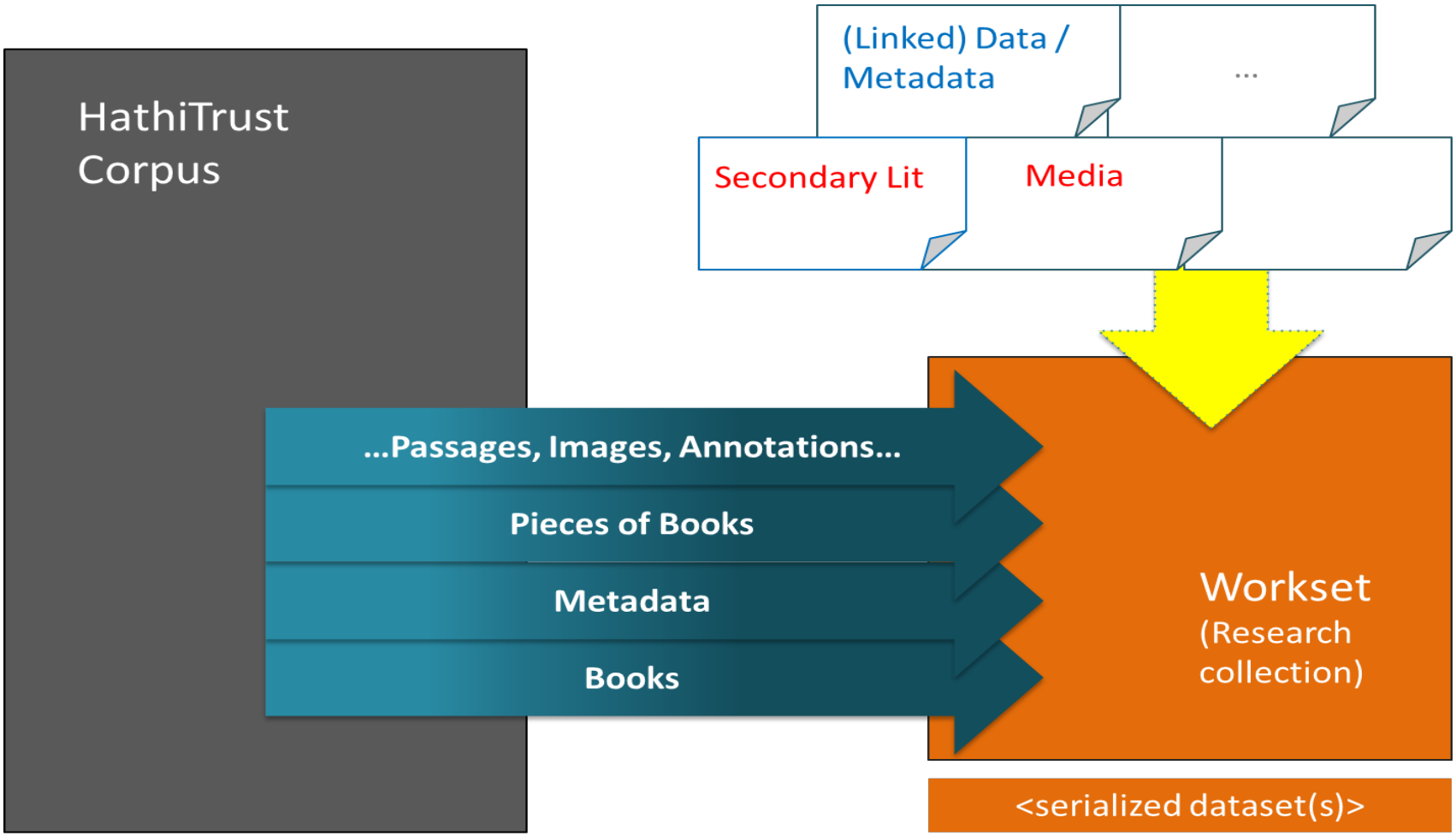
What is a Workset (in context of HTRC)?

- A workset is an aggregation brought together for the purpose of analysis, i.e., to facilitate inquiry.
- Worksets are conceptual and need to be expressible in a variety of ways
- A workset encapsulates the specific materials that share specific attributes or satisfy some set of criteria.

- May be large, e.g., tens of thousands of items.
- Can be constructed by machine as well as human agents.
- Attributes and criteria not always bibliographic
- Items aggregated may be more granular than a volume

Why Worksets?

- The result of a first-level, rough filter
- Better scale for intensive analytics
- Provides essential scope for certain analytics
- 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


HathiTrust
Corpus
(Linked) Data /
Metadata
...
Secondary Lit
Media
...Passages, Images, Annotations...
Pieces of Books
Metadata
Books
Workset
(Research
collection)
<serialized dataset(s)>

Scope

Workset Creation for Scholarly Analysis
Prototyping Project

Collection analysis, data modeling and prototype tools & services to facilitate workset creation

- Principal investigators:
  J. Stephen Downie, Tim Cole, Beth Plale
- Funded by a grant from the Andrew W. Mellon Foundation
- 1 July 2013 - 30 June 2015

- Will feature 4 $40K sub-awards for prototyping/demonstration projects illustrating how worksets from HT DL can be created and used and can be useful for scholarly analysis
  - Methods & tools for metadata enrichment, including with links
  - Analytical services over full text useful for defining worksets

Key research questions for WCSA project

- Can we formalize the notion of collections and worksets in the HTRC context?
- What are the attributes that define and describe a workset in the context of HTRC?
- How can we balance rigor with extensibility & flexibility?
- What roles do data, metadata, annotations, tags, feature sets, and so on, play in the conception, creation, use and reuse of collections and worksets?

Can we demonstrate the utility & practicality of worksets for HTRC?

WCSA Timeline

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

USER NEEDS & REQUIREMENTS

Harriett Green, English and Digital Humanities Librarian

- Preliminary results
- An early deliverable of WCSA Project

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

- Report prepared for the HTRC in 2011 by UIUC researchers at GSLIS's Center for Informatics Research in Science and Scholarship (CIRSS)
- Interviewed researchers who were awarded Google Digital Humanities Research Awards on research needs
- Findings for scholarly requirements included improved metadata, accurate OCR, data curation
- Report available to download at http://www.hathitrust.org/htrc

Feedback from UnCamp 2013

My work-set should contain…

- Volumes pertaining to Japan / in Japanese
- Music scores or notation extracted from HT volumes
- Volumes in HT similar to TCP-ECCO novels

General Needs:

19th c. English-language novels by female authors

- User-friendly interfaces
- Documentation on the portal
- Avenues for community input in HTRC portal development

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

- Conducted at DH 2013, JCDL 2013, and HTRC Uncamp conferences in summer and fall 2013
- **Goal:** To understand practices of humanities researchers using digital collections, especially in the context of large-scale text corpora
- Survey instrument queried users about their experiential practices of organizing datasets

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, history
- Institutions:
  - Academic institutions in Great Britain, Singapore, Germany, France, and United States

Study Design

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

Analysis

Methodology: Qualitative content analysis of user responses

- A "directed" approach based on inductive reasoning to condense raw data (transcriptions of audiorecordings of interviews and focus groups) into categories and themes

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

Coding (still ongoing)

Coding manual consisting of category names, rules for assigning codes, and examples:

- Challenges — access rights
- Challenges — OCR quality
- Collections — comprehensiveness
- Objects — data
- 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.

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 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."

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.

Use Case 1: Gender

- Scholar wants to compare works by gender, based on the Library of Congress headings
- This information is in the metadata, but hard to text mine
- Questions:
  - How can I track gender of authored texts across time?
  - What correlations are there between gender of the author and sentiment analysis of the text?
  - How people and characters of different genders are treated in books over time?

Use Case 2: Serials

A scholar wants to find a series of an author's works that were originally serialized across several issues or volumes of a periodical.

- Serials vs. volumes as manifestations of works
- Map the pages for content
- Might be able to investigate questions as:
  - What was the original instantiation of the work in serialized form?
  - How can I text mine for sentiment and themes across the serialized texts?

Use Case 3: Images

Scholar wants to find texts of Victorian travel narratives and the images depicted in them.

Investigate questions such as:

- What are patterns/themes of images depicted in Victorian England travel narratives?
- What is the frequency of images in travel narratives?

Use Case 4: Dialogue in Texts

Scholar wants to identify conversational dialogue between characters in novels.

- Requires OCR that detects boundaries: can we detect quote marks and signal words for dialogue?
- Create a training set of curated texts (i.e., TCP texts) matched with HTRC texts, apply detection algorithm
- Enable questions such as:
  - How are characters connected across the narrative—who interacts most frequently?
  - What would sentiment analysis or topic modeling reveal about the dialogue in comparative novels of the genre?

User Needs for Worksets

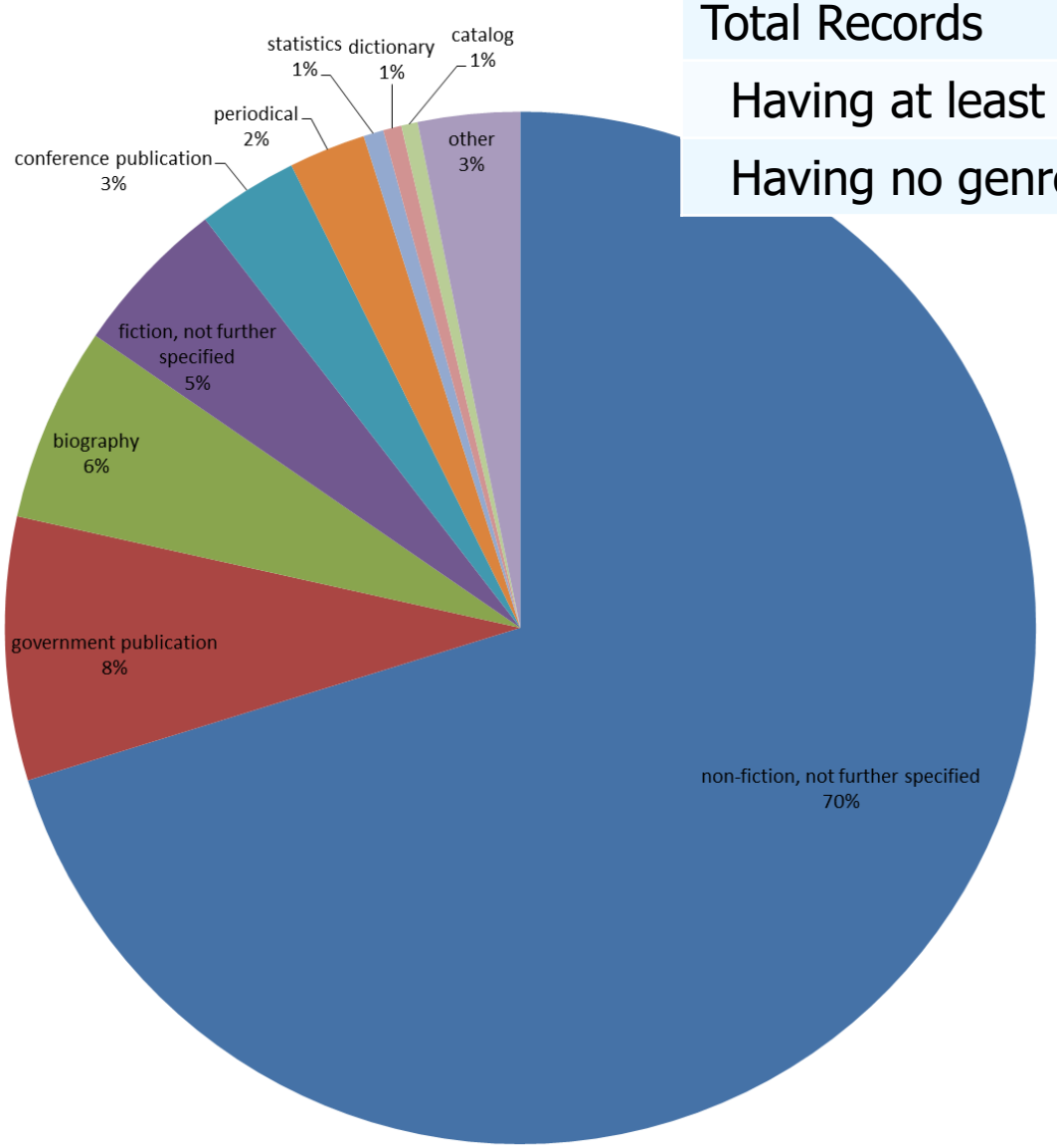
Comments from interviews/focus groups:

"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?"

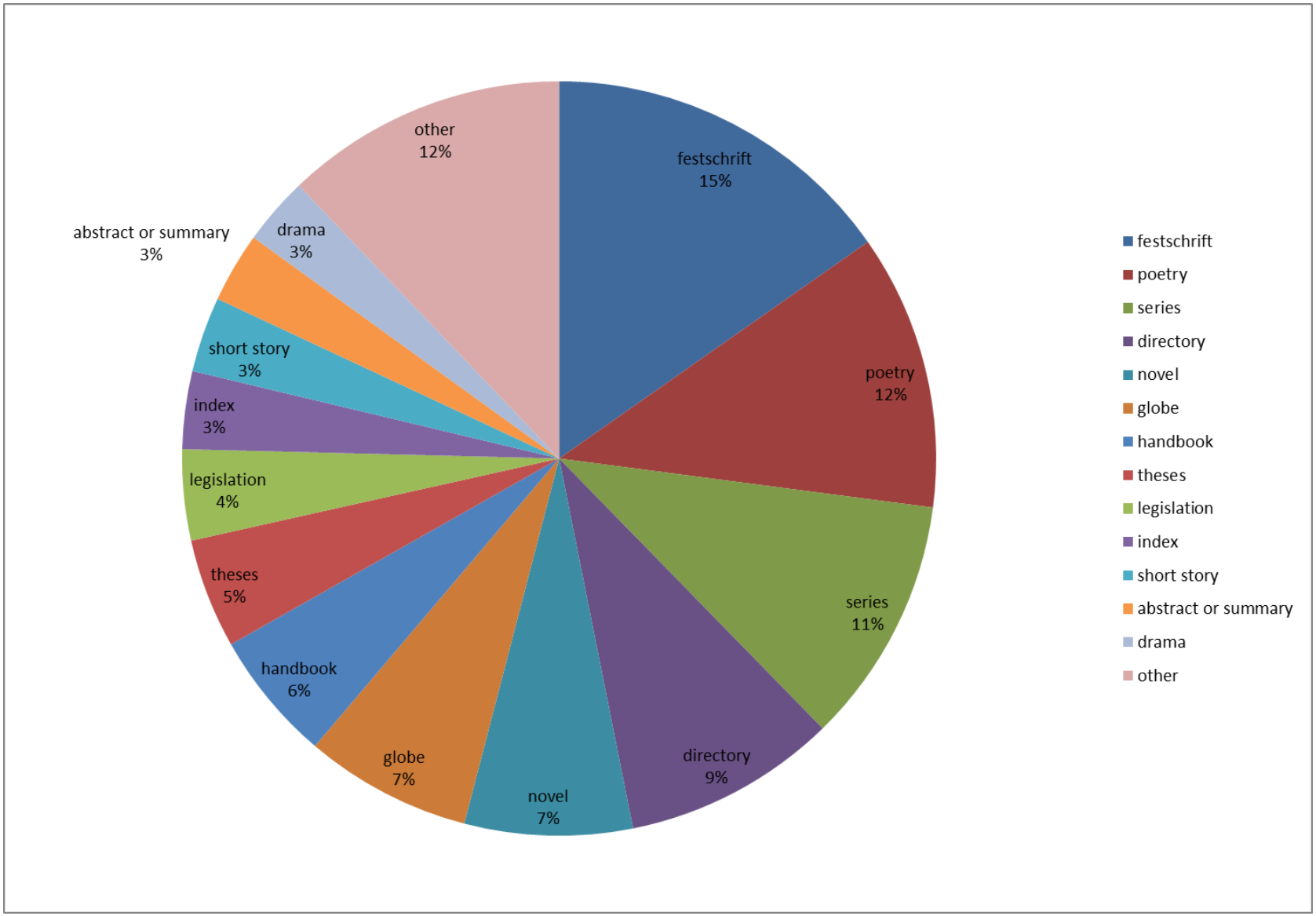
How useful is existing metadata for creating worksets?

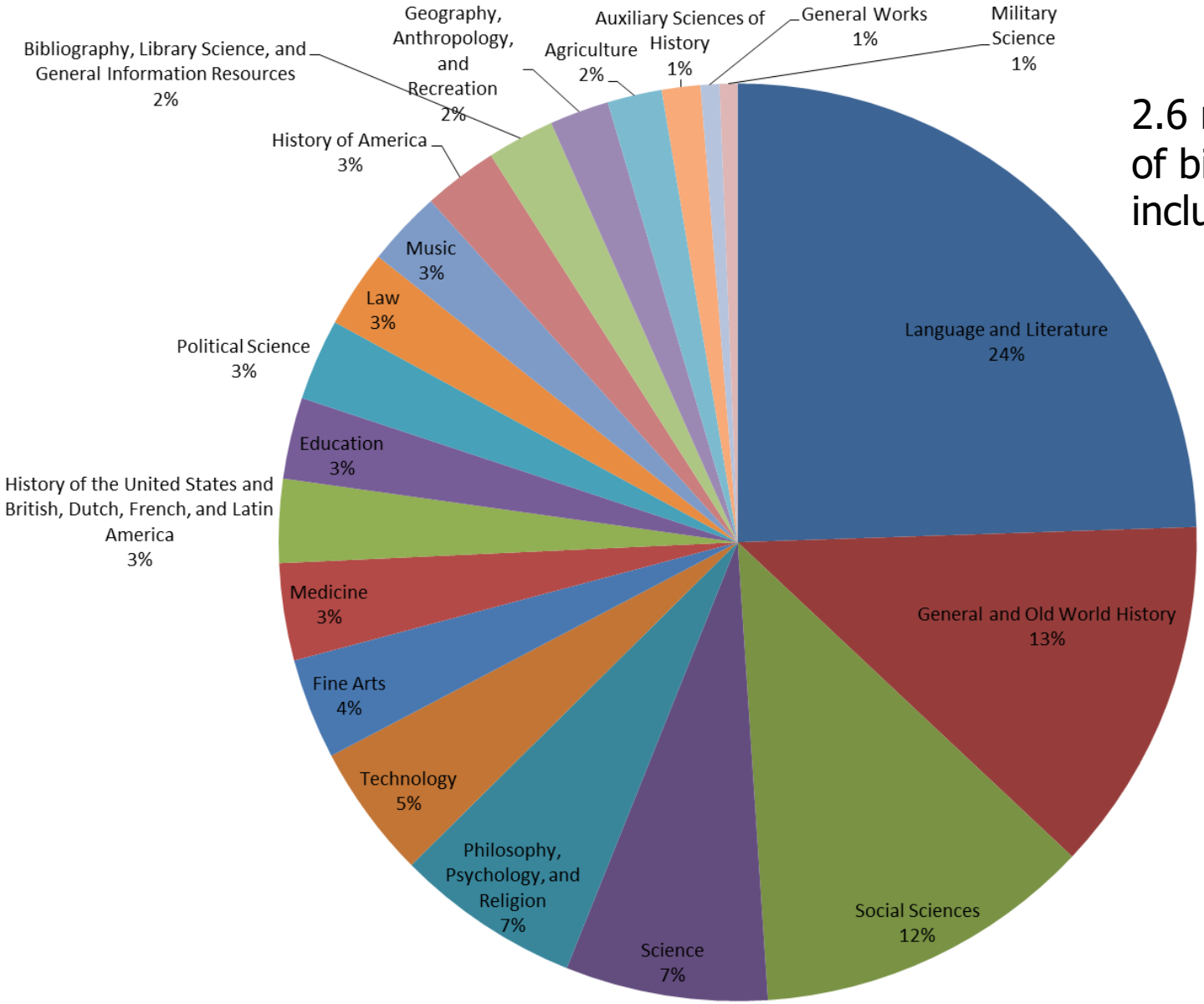
- HT metadata is bibliographic
- Built from MARC records provided by members & OCLC
- Good, consistent quality for author / title / pub info
- Subject less extensive, less consistent
- Genre more hit and miss
- Author gender not present in MARC bib records
  is present in some MARC authority records
- MARC records provided for serials are about the serial
  not about the contents of the serial
- No visibility over internal elements (e.g., images, embeded
  language / genre, dialog, ...) of digitized volumes



Total Records	6.1 million	
Having at least 1 genre	5.2 million	85%
Having no genre	0.9 million	15%

Breakdown of other category, incl. fiction





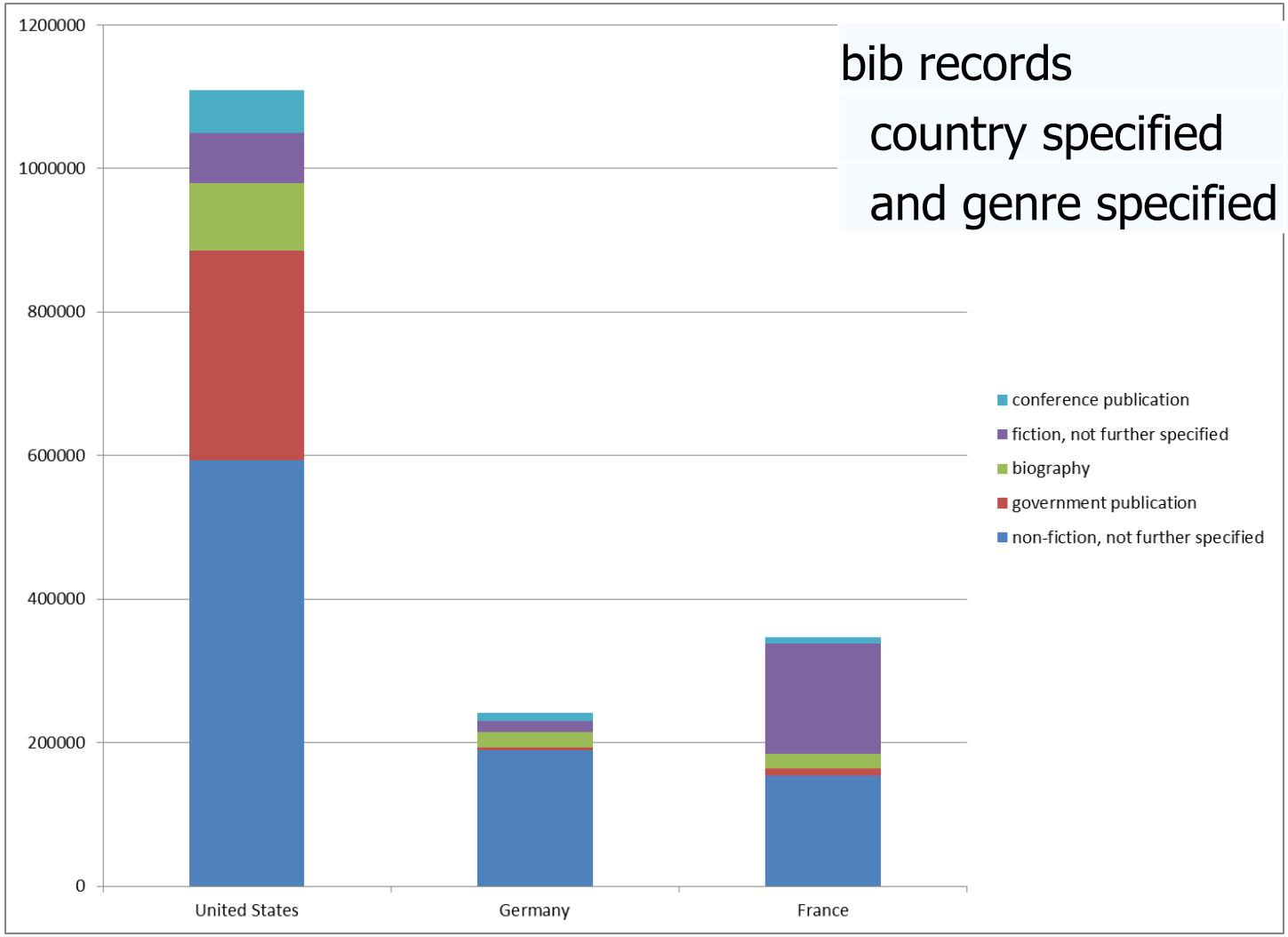
2.6 million (43%) of bib records include LC Class no.

Timothy W. Cole *(t-cole3@uiuc.edu)*
University of Illinois at UC

Not all genres equally described

Volumes identified as fiction	270837	
at least one subject	70706	26%
no subject	200131	74%
subectGeographic	25491	9%
subjectTemporal	12536	5%
subjectTopic	61788	23%
subjectName	13412	5%

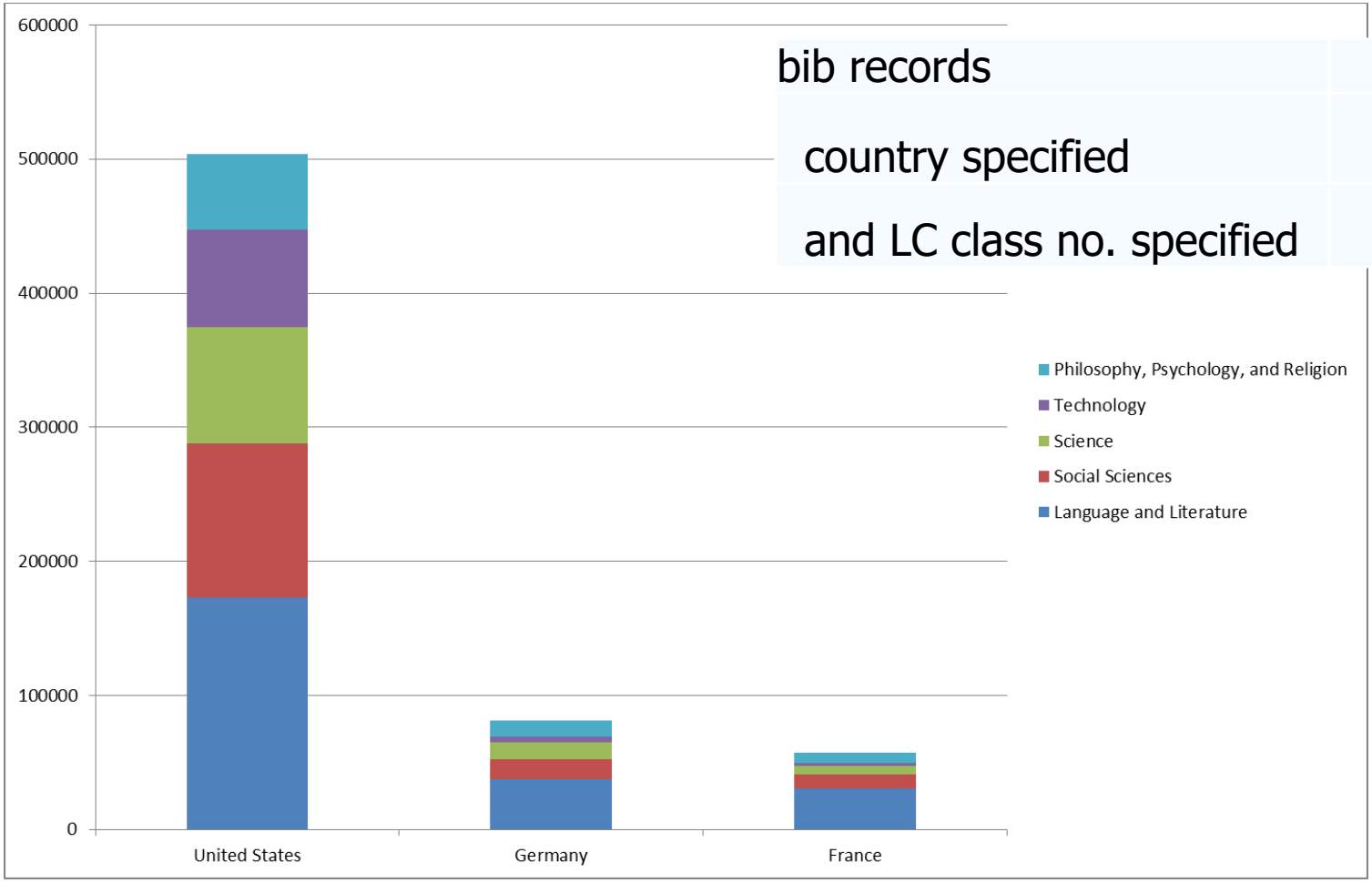
Top genres by country of publication



bib records	6067835
country specified	5361473
and genre specified	4776973

Timothy W. Cole *(t-cole3@uiuc.edu)*
University of Illinois at UC

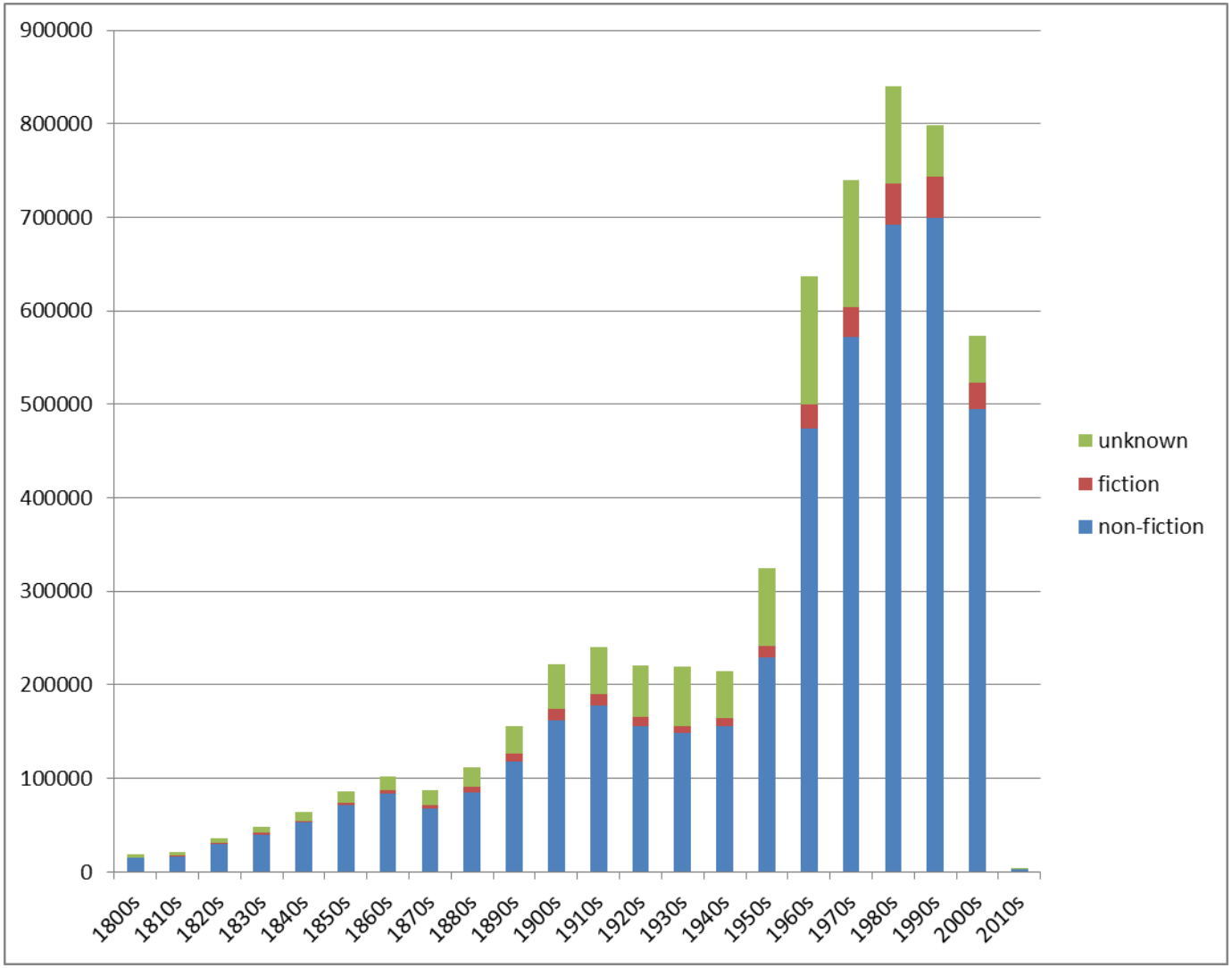
Top subjects by country of publication



bib records	6067835
country specified	5361473
and LC class no. specified	2354094

Timothy W. Cole *(t-cole3@uiuc.edu)*
University of Illinois at UC

Fiction as proportion of publications by decade



Timothy W. Cole *(t-cole3@uiuc.edu)*
University of Illinois at UC

Opportunities

- Computed attributes
  - Author age at time of publication
  - FRBR relationships
- Add attributes not included in bibliographic records
  - Author gender
  - Author nationality
- Improve completeness & accuracy of bib records
- Describe internal components of volumes

More about RFP

- 4 awards to teams of scholars, librarians & developers
  - $40,000 each
  - Period of performance 16 April 2014 – 15 Jan 2015
  - UIUC will supply a testbed of ~250,000 representative volumes; additional volumes (digitized from public domain) available
  - UIUC will collaborate, provide access to HTRC cluster, ...
  - Deliverables: final report; open source software

- Schedule:
  - Letters of Intent Due (preferred): 16 December 2013
  - Final Proposals Due: 13 January 2014
  - Shortlist Meeting Invitations Issued: 20 January 2014
  - Shortlist Meeting: 20 February 2014
  - Award Notification: No later than 15 March 2014

Questions?

Timothy Cole
Mathematics and Digital Services Librarian, UIUC
t-cole3@illinois.edu

Harriett Green
English and Digital Humanities Librarian, UIUC
green19@illinois.edu
Twitter: @greenharr

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-producibility implications for such highly individualized and complex digital objects