# Digital Humanities At Scale: Hathi Trust Research Center

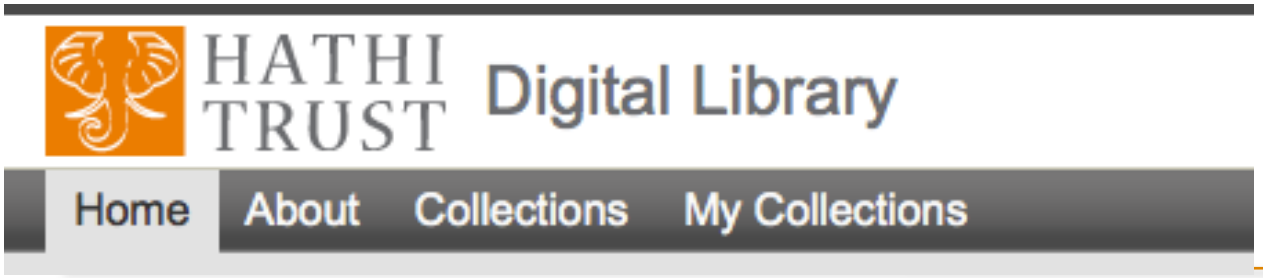
Notre Dame digital humanities, May 7, 2013

Beth Plale, Indiana University


#HTRC #HathiTrust

# HTRC Mission

- Public research arm of the HathiTrust
- Help researchers world-wide to accomplish tera-scale text data-mining and analysis
  - Develop cutting-edge software tools for processing, analyzing text
  - Develop cyberinfrastructure to enable HPC access to the HathiTrust Digital Library
- Established: July, 2011
- Collaborative center: Indiana University & University of Illinois




## Currently Digitized

- 10,727,868 total volumes
- 5,616,955 book titles
- 278,428 serial titles
- 3,754,753,800 pages
- 481 terabytes
- 127 miles
- 8,716 tons
- 3,375,498 volumes(~31% of total) in the public domain

View visualizations of HathiTrust call numbers, languages, and dates

→ HathiTrust is large corpus providing opportunity for new forms of computation investigation.

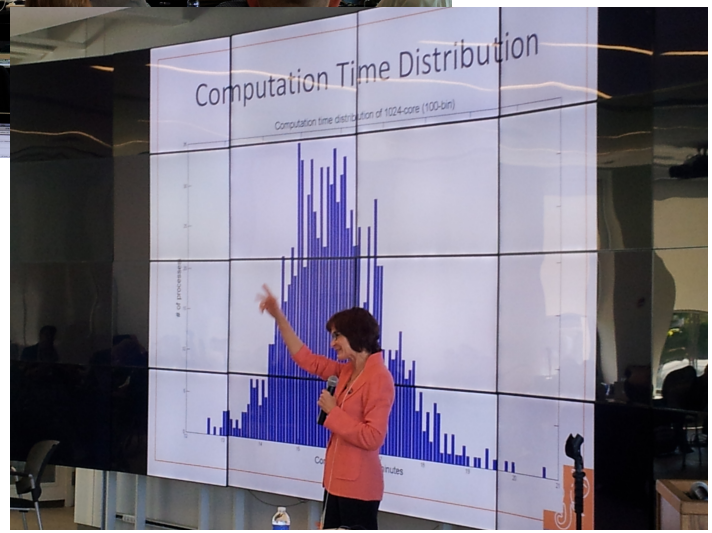
→ The bigger the data, the less able we are to move it to a researcher's desktop machine

→ Future research on large collections will require ***computation moves to the data, not vice versa***

# HTRC Next Steps

- Phase 2 availability of resource 31 March 2013
- Thanks to:

ALFRED P. SLOAN FOUNDATION




Photos from HTRC UnCamp 9.10.12
at Indiana University

# HTRC Non-Consumptive Research Paradigm

- *No action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from collection of copyrighted works to reassemble pages from collection.*
- Definition disallows collusion between users, or accumulation of material over time. Differentiates human researcher from proxy which is not a user. Users are human beings.

# Initial Requirements Gathering: 2010-11

GOOGLE DIGITAL HUMANITIES AWARDS RECIPIENT INTERVIEWS REPORT
PREPARED FOR THE HATHITRUST RESEARCH CENTER

VIRGIL E. VARVEL JR.
ANDREA THOMER
CENTER FOR INFORMATICS RESEARCH IN SCIENCE AND SCHOLARSHIP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Fall 2011

# The study

- John Unsworth invited all 22 researchers with Google Digital Humanities Research Awards to participate in study
- Interviews were conducted via telephone, Skype®, or face-to-face, and all were audio recorded. All participants agreed to IRB permission statement via email.
- A semi-structured interview protocol was developed with input from HTRC to elicit responses from participants on primary goals of project.

# Select findings

- Optical Character Recognition
  - Improve OCR quality where possible
  - Enhance scanned image views for OCR reference and correction
  - Metadata should expose the quality of OCR
- Need better, granular metadata about languages (human correction preferred)
- Need Bibliographic records in useable form

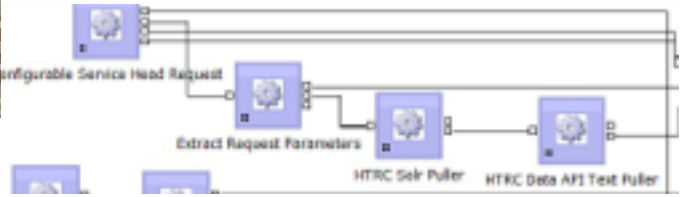
# Goals for HTRC

- Provide a persistent and sustainable structure to enable original and cutting edge research.
  - Leverage data storage and computational infrastructure at Indiana & Illinois
  - Stimulate community development of new functionality and tools
  - Use tools to enable discoveries that would not be possible without the HTRC
- Enable scholars to fully utilize content of HathiTrust Library while preventing intellectual property misuse within U.S. copyright law.
  - Provision secure computational and data environment for scholars to perform research using HathiTrust Digital Library.

# New Questions

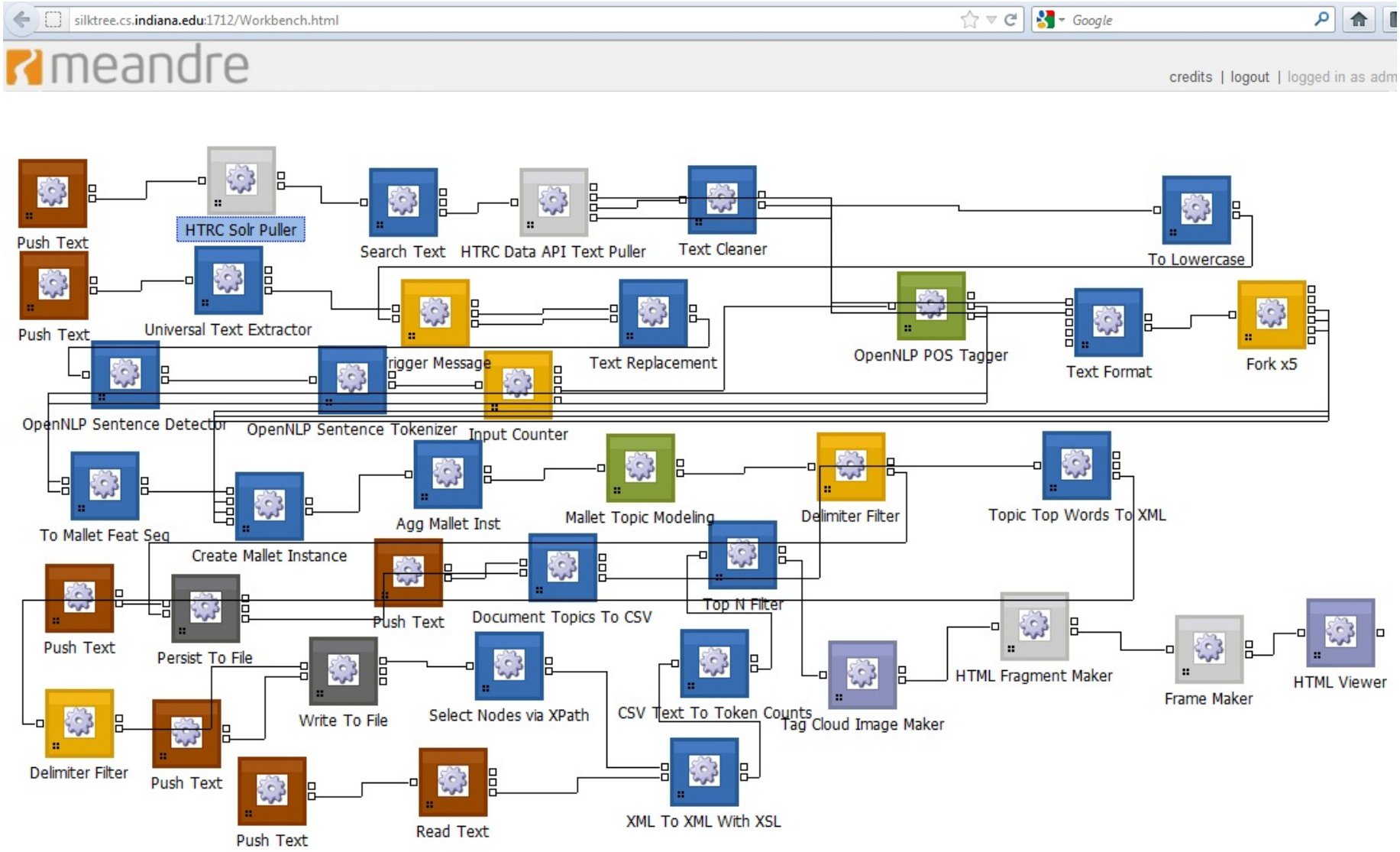
Identify all 18th century published books in HathiTrust corpus, and apply topic modeling to create consistent overall subject metadata

- Ted Underwood *et al*., University of Illinios



# Topic Modeling


- Can answer more complex or nuanced questions
  - What are the primary themes of an author?
  - What are the primary themes of a research domain?
  - When did a new topic enter a research domain?
- Provides more data than word counts
  - 100s of topics can be extracted.
  - Underlying data (topics, volume, and page) is available




Topic Modeling workflow

# Major Theme for an Author

## Charles Dickens

- 195 volumes in the HTRC non-Google collection
- 100 topics generated

# Themes for Authors

- Two topics with identical centralities but separate themes


## Exemplar HTRC Research:

*The task of cleaning and enriching large collections: what aspects can we share?*


UIUC English Dept.:

***Ted Underwood***

Jordan Sellers

Mike Black

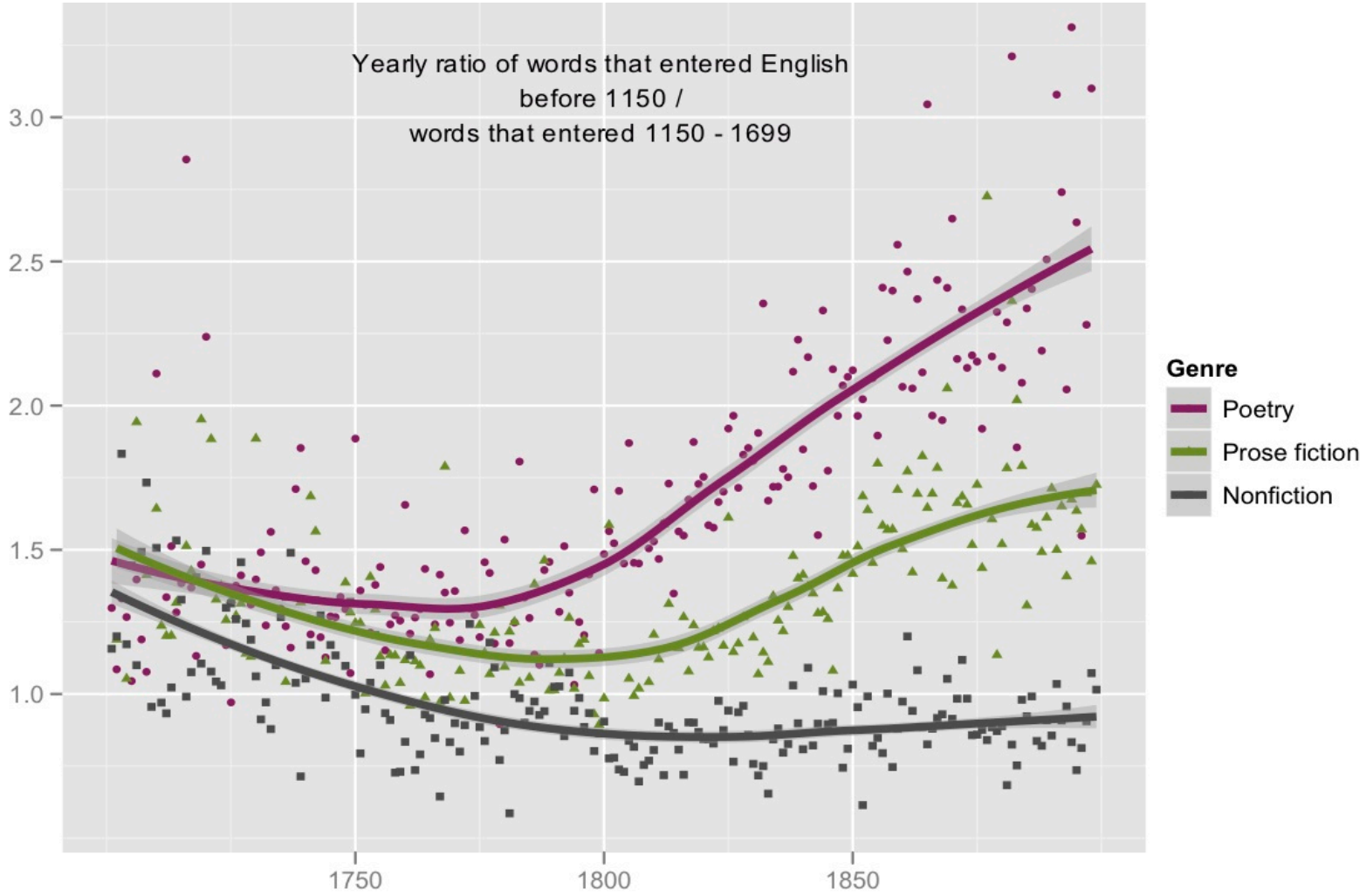
UIUC Library: Harriett Green

I3: Loretta Auvil, Boris Capitanu

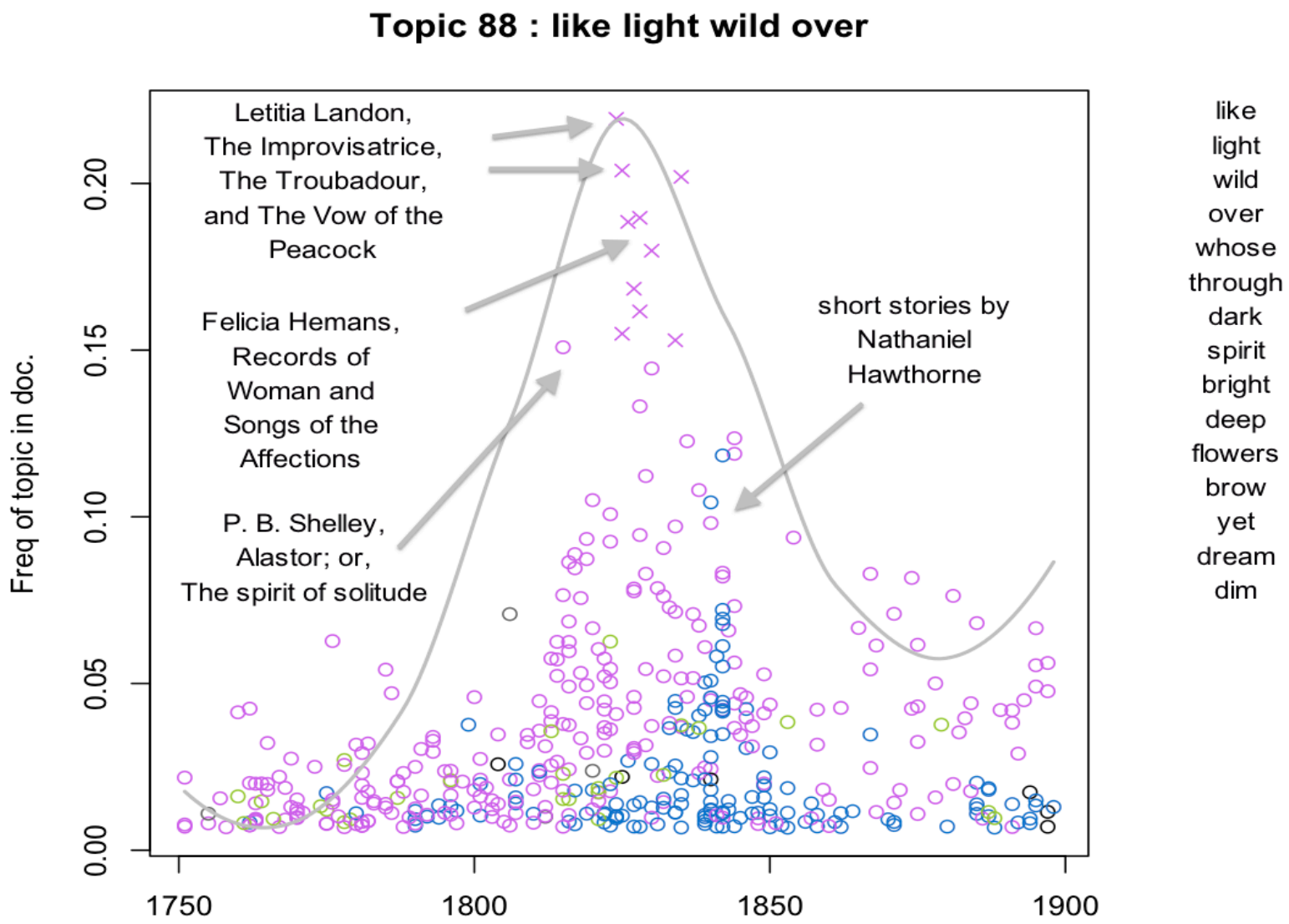


*Supported by: The Andrew W. Mellon Foundation*

# Yearly values of a ratio between two wordlists in three different genres. 4,275 volumes. 1700-1899.



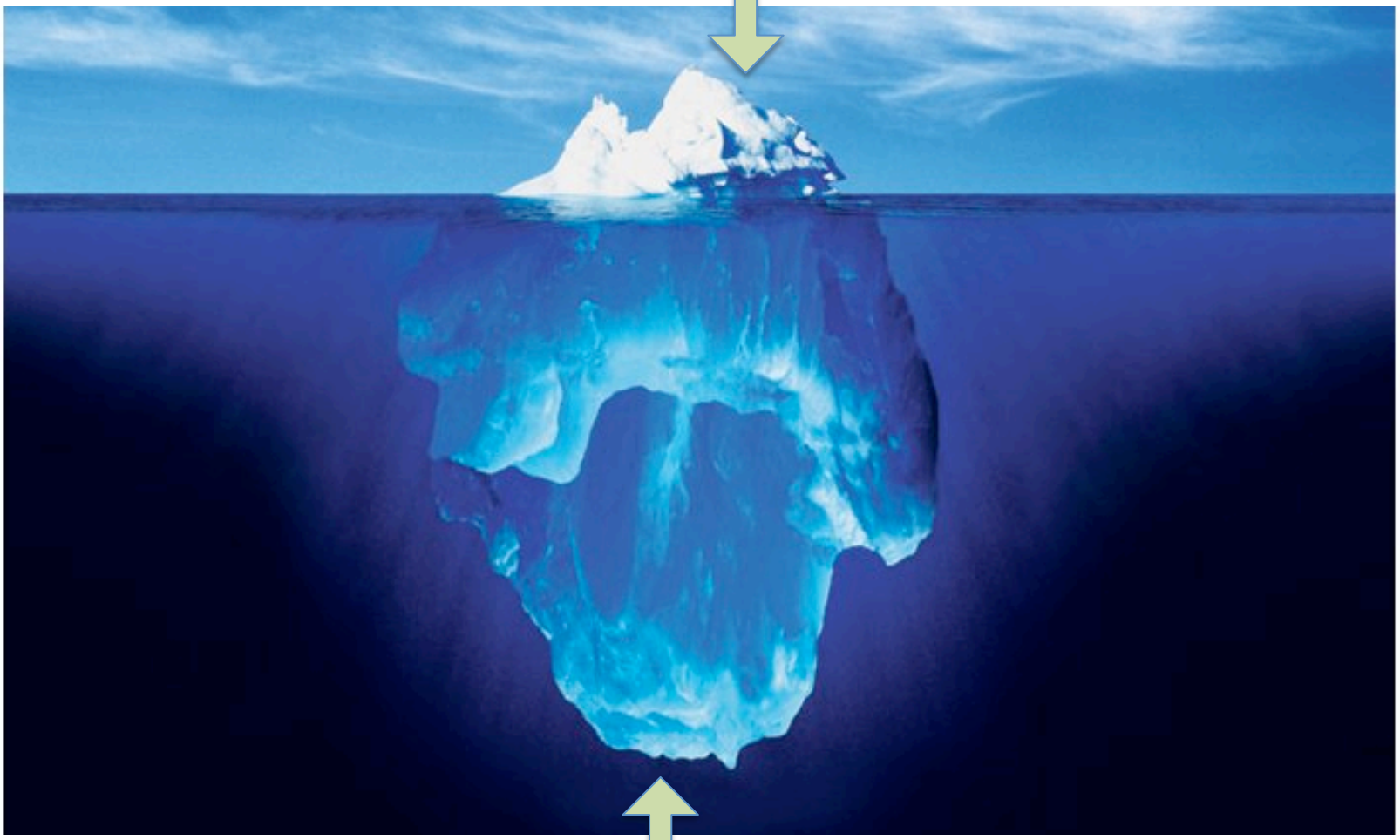

Underwood et al. Research




Blue/fic, purple/poe, green/drama, black/bio, brown/nonfic, triangle/letters or orations.

Underwood et al. Research

analyzing the data



cleaning the data

Underwood et al. Research

# Cleaning the data

1. Clean up the OCR / assess error.
2. Identify parts of a volume (e.g., articles in a serial, poetry/prose).
3. Remove library bookplates and running headers — after using them for (3).

# Cleaning/enriching the metadata

1. Discard duplicate volumes / select early editions?
2. Add metadata that you need for interpretive purposes, like
   — gender (see Ben Schmidt's technique),
   — genre.

Underwood et al. Research

# Things we could share

period lexicons / variant spellings

gazetteers of proper nouns

OCR correction rules for a period

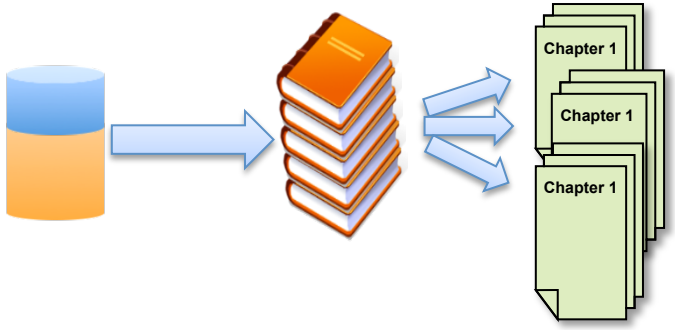
document segmentation and/or cleaned and segmented text

ferberization

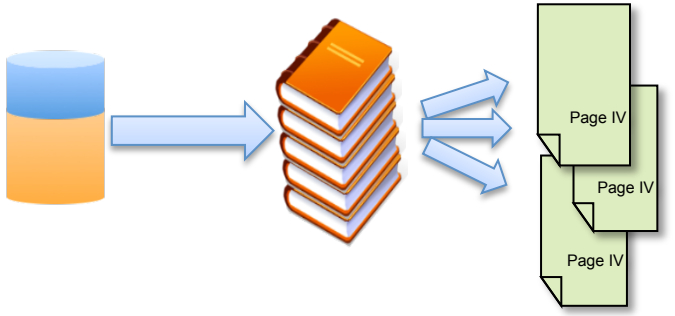
cleaned / enriched metadata

… and of course, share code to do all of above

# Corpus Usage Patterns




Access by chapter

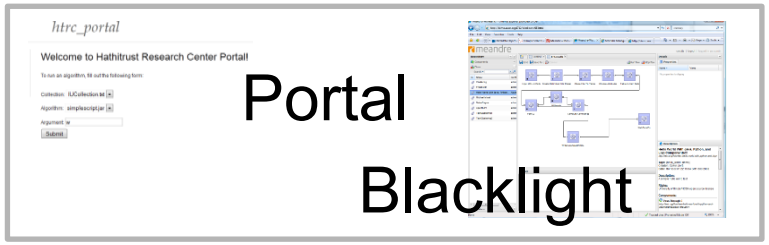



Access by page


Access by special contents (table of contents, index, glossary)

# HTRC architecture


- Philosophy:  computation moves to data
- Web services architecture and protocols
- Registry of services and algorithms
- Solr full text indexes
- noSQL store as volume store
- openID authentication
- Portal front-end, programmatic access
- SEASR mining algos




RESEARCH CENTER

CILogon

Agent framework

SEASR analytics service

WSO2 registry
services, collections, data capsule images

HTRC Data API v0.1

Access control (e.g. Grouper)

Agent instance

Agent instance

Agent instance

Agent instance

Solr index

Programmatic access e.g.,

PROJECT bamboo

Task deployment

Meandre Orchestration

Non-consumptive Data capsules

Volume store (Cassandra)

NCSA local resources

Future Grid

Big Red II

NSF XSEDE

Page/volume tree (file system)

rsync

HathiTrust corpus

HATHI TRUST

University of Michigan

# Algorithms

- Computational analysis is accomplished through algorithms
  - An algorithm carries out one coherent analysis task: sort list of words, compute word frequency for text
- Researcher's computational analysis often requires running sequence of algorithms. Important distinction for implementing non-consumptive research is "who owns the algorithm"?

# Infrastructure for computational analysis

- When needing to support computation over 10+M volume corpus, algorithms must be co-located with data.
- That is, algorithms must be located where repository is located, and not on user's desktop.
- When computational analysis is to be non-consumptive, likely one location for the data.

# Who owns algorithm?

- HTRC owns the algorithms,
  - use Software Environment for Advancement of Scholarly Research (SEASR) suite of algorithms
  - we are examining security requirements of users, algorithms, and data

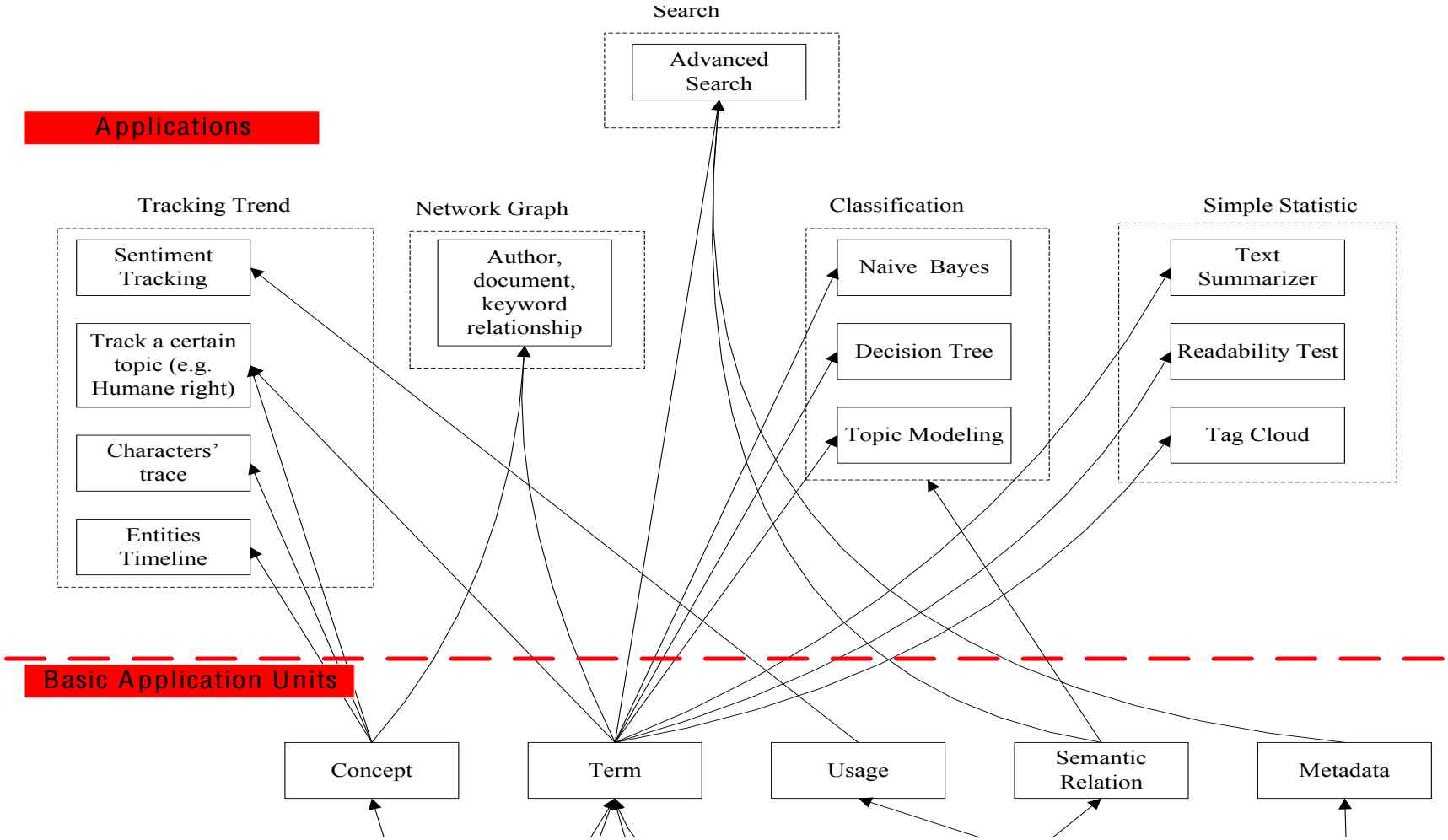
# User owns and submits their algorithms

- HTRC-Sloan-Cloud - principle of “trust but verify”. Informatics-savvy humanities scholar is given freedom to experiment with new algorithms on protected information, but technological mechanisms in place to prevent undesirable behavior (leakage.)

# HTRC-Sloan-Cloud

- Implements non-consumptive
- Openness – users not limited to using known set of algorithms
- Efficiency – Not possible to analyze algorithms for conformance prior to running
- Low cost and scale – Run at large-scale and low cost to scholarly community of users
- Long term value –adoption for other purposes

# Categories of algorithms. Can fair use be determined based on categorization of algorithm? Or is all computational use fair use?



# Algo results fair use?

- Center supplied
  - Easier because we know category of algorithm
- User supplied
  - HTRC is not examining code, so open question

# Parting philosophy

- Finally, results of computational research that conforms to restrictions of non-consumptive research must belong to researcher

# HTRC Phase II : Objectives

- Outreach:  plan  and budget for '13-'14 AY
- Software development:  Streamline development effort. Priority on:
  - ***User-driven requirements: track, prioritize***
  - ***Bugs***
  - ***Simplification/ease of management***
  - ***HTRC Sloan Cloud for non-consumptive research***
- Improved funding efforts – stronger position
- Improved reporting / tracking

# HTRC Tech Stack Deployment Timeline

## Deliver: Mar 31, 2013

- **Sandbox stack (resides at UIUC):** non-google corpus (250,000 volumes), open access.
- **Production stack (resides at IU):** v0.5 in place. Uses Oauth security. Public domain corpus. Shares Cassandra/Solr with dev stack. Minimal compute resources available.
- **Development stack (resides at IU):** shares Cassandra/Solr with prod stack. **Supports v0.1 of HTRC Sloan Cloud for non-consumptive support**

## Deliver: Jun 30, 2013

- **Sandbox stack (at UIUC):** v1.0 stack but against non-google corpus
- **Production stack (at IU):** v1.0 reflects extensive testing. Oauth for security. Public domain corpus. Share Cassandra/Solr with dev stack. Support for parallel execution.
- **Development stack (at IU):** share Cassandra/Solr with prod stack. New services. V0.2 of Sloan non-consumptive support. Begin dev for InCommon and auditing.

## Deliver: Sep 30, 2013

- **Sandbox stack (at UIUC):** v1.5; against non-google corpus
- **Production stack (at IU)**: v1.5. Supports inCommon in anticipation of copyright works. Public domain corpus. Separate Cassandra/Solr; public domain corpus
- **Development stack (at IU):** InCommon, auditing, and v1.0 of Sloan non-consumptive support. Security audit on development stack; verify ready for copyright materials

## Deliver: Nov 30, 2013

- **Sandbox stack:** retire (?)
- **Production stack (at UIUC or IU)**: v2.0. Supports inCommon in anticipation of copyright works. Public domain corpus. Separate Cassandra and Solr for public domain corpus.
- **Development stack (at IU or UIUC):** dev stack ready for copyright materials.



RESEARCH CENTER

# The Workset

- Workset Defn: *set of pointers to all or part of any number of items in the HT corpus and external to the corpus*
- HTRC v1.0 has crude notion of collection as list of volume IDs.
- HT has "collection builder", collection built manually then saved. People in text analytics need to gather many objects (10,000), can't be built manually (augment workset by learning from hand-built set).
- Reimagine what objects are:
  - Could be pictures on a page. Deconstructing the page, the volume. Notions of page, chapter. Ability to point at, and move around. Aggregations of things within works.
  - Points to 'things' that are also outside HTRC: e.g. sentiment label stored in semantic web. This workset (similar to research object) is then passed in for computation.
- Provenance of analysis process for reproducibility

# Add value to corpus

- Services that add value:
  - Gender detector: run on 10 M volumes. "On p. 52 detected a female voice". Return page number and label. Or gender of author.
  - Mining metadata of a collection; used to describe a collection more fully. Provides context information about collections.
  - Error correction in the OCR. Adding classifiers to metadata.
  - Run off-line (at night)
  - Is there corpus augmentation we could undertake to prototype (of high value)? Would need to be meaningful on whole corpus versus meaningful on portion of corpus.

# How to Engage

- Uncamp 2013, Sept 13-14, Urbana, Illinois
- AY '13-14 is community outreach phase of HTRC: looking for friendly community of researchers: build partnership; get code running on HTRC; help with parallelization
- *Workset Creation for Scholarly Analysis: Prototyping Project,* Mellon proposal (pending) – community funded projects with direct impact on HT corpus


# Thank You


- This presentation was made possible with content provided by many HTRC colleagues John Unsworth, J. Stephen Downie, Robert McDonald, Beth Sandore, Yiming Sun, Guangchen Ruan, Loretta Auvil, Kirk Hess, and many others…
- The HTRC Non-Consumptive Research Grant is graciously funded by the Alfred P. Sloan Foundation
- IU D2I-PTI is graciously funded by The Lilly Endowment, Inc.


- HTRC - http://www.hathitrust.org/htrc
- IU D2I Center - http://d2i.indiana.edu/
- UIUC GSLIS - http://www.lis.illinois.edu/

# Contact Information

- Beth Plale, IU,
  - plale@indiana.edu
- Technical
  - Yiming Sun, Chief Architect, yimsun@indiana.edu
- Requests for capability, interest
  - Miao Chen, HTRC Asst. Director of Education and Outreach, miaochen@indiana.edu