RESEARCH CENTER

RICE
Unconventional Wisdom

Baylor
College of
Medicine

ILLINOIS

Northeastern University

HT+BW
HathiTrust + Bookworm

Loretta Auvil
University of Illinois
DPLAFest 2015
April 18, 2015

NATIONAL ENDOWMENT FOR THE Humanities

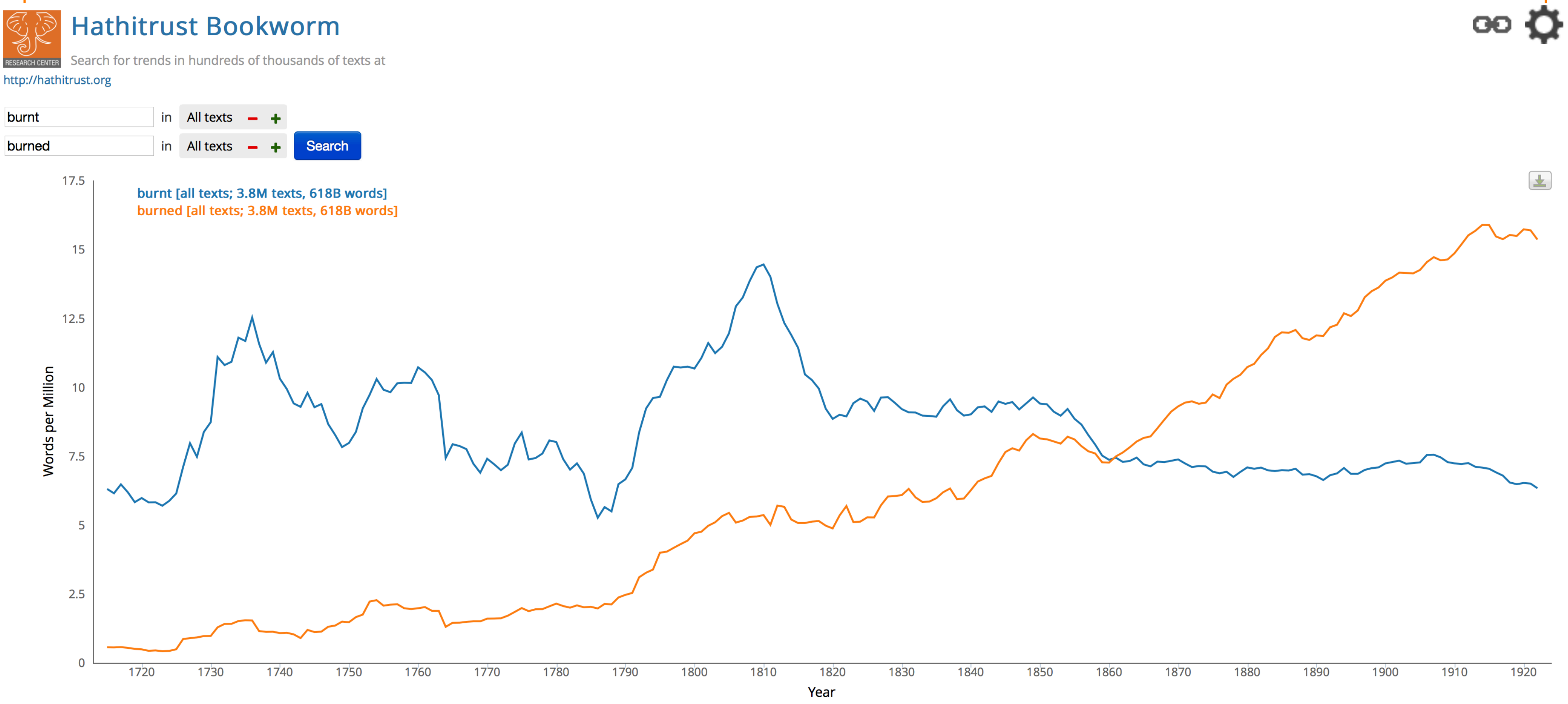
NEH Implementation Grant

- Exploring the Billions and Billions of Words in the HathiTrust Corpus with Bookworm
- Team Members
  - J. Stephen Downie, University of Illinois at Urbana-Champaign
  - Erez Lieberman Aiden, Baylor College of Medicine
  - Benjamin Schmidt, Northeastern University
  - Robert McDonald, Indiana University
  - Loretta Auvil, University of Illinois at Urbana-Champaign
  - Sayan Bhattacharyya, University of Illinois at Urbana-Champaign
  - Muhammad Shamim, Baylor College of Medicine
  - Peter Organisciak, University of Illinois at Urbana-Champaign
- Former Team Members
  - Colleen Fallaw, University of Illinois at Urbana-Champaign
  - Matt Nicklay, Baylor College of Medicine

HT+BW Project

- HathiTrust (HT)
  - Text data
  - Meta data
- Bookworm (BW)
  - Tool that visualizes language usage trends in repositories of digitized texts in a simple and powerful way

Irregular Verbs Changing Over Time



Hathitrust Bookworm

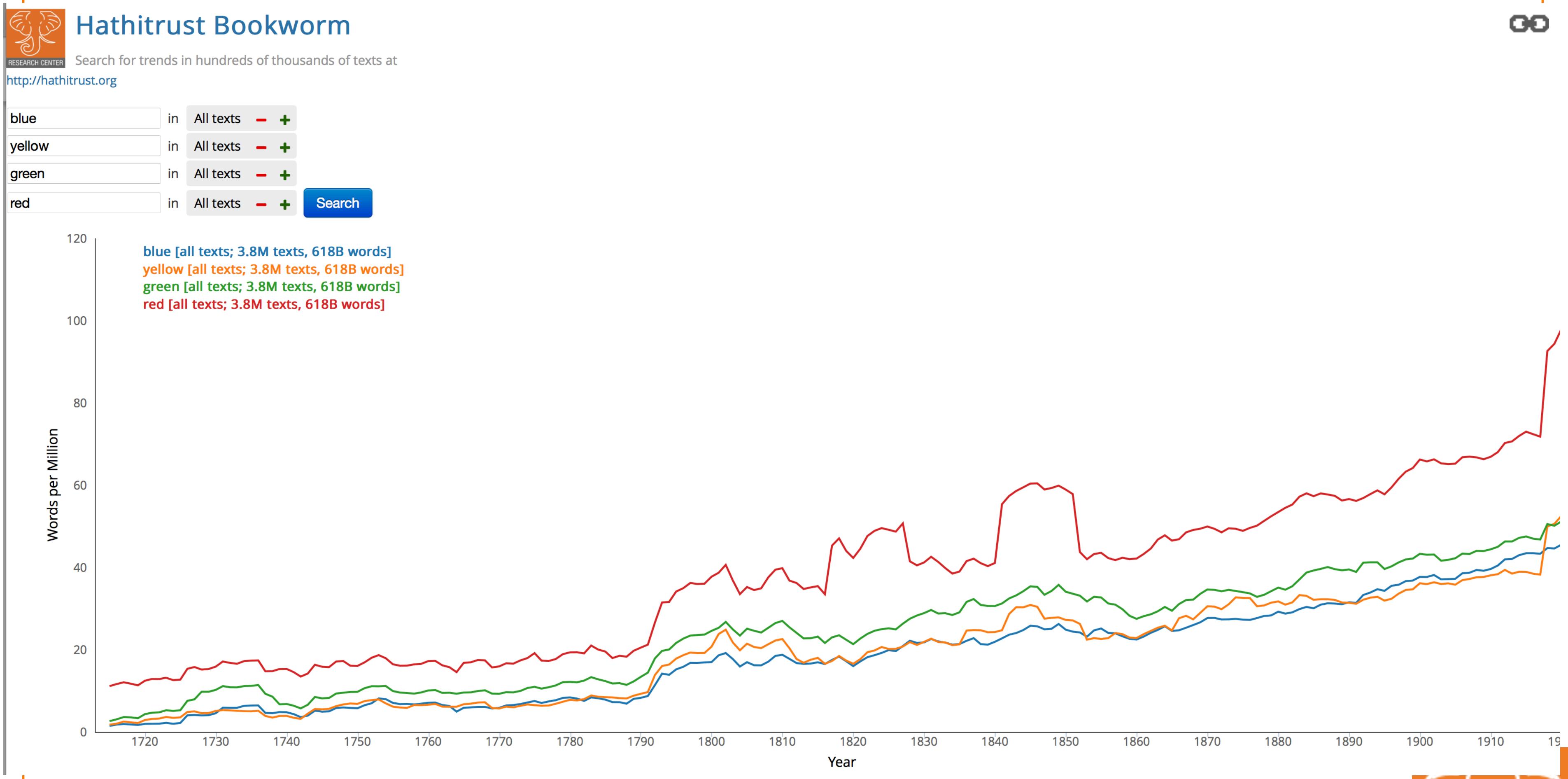
Search for trends in hundreds of thousands of texts at http://hathitrust.org

burnt in All texts – +

burned in All texts – + Search

Trend of burnt vs burned

Tracking Colors



Hathitrust Bookworm

Search for trends in hundreds of thousands of texts at http://hathitrust.org

blue in All texts – +

yellow in All texts – +

green in All texts – +

red in All texts – + Search

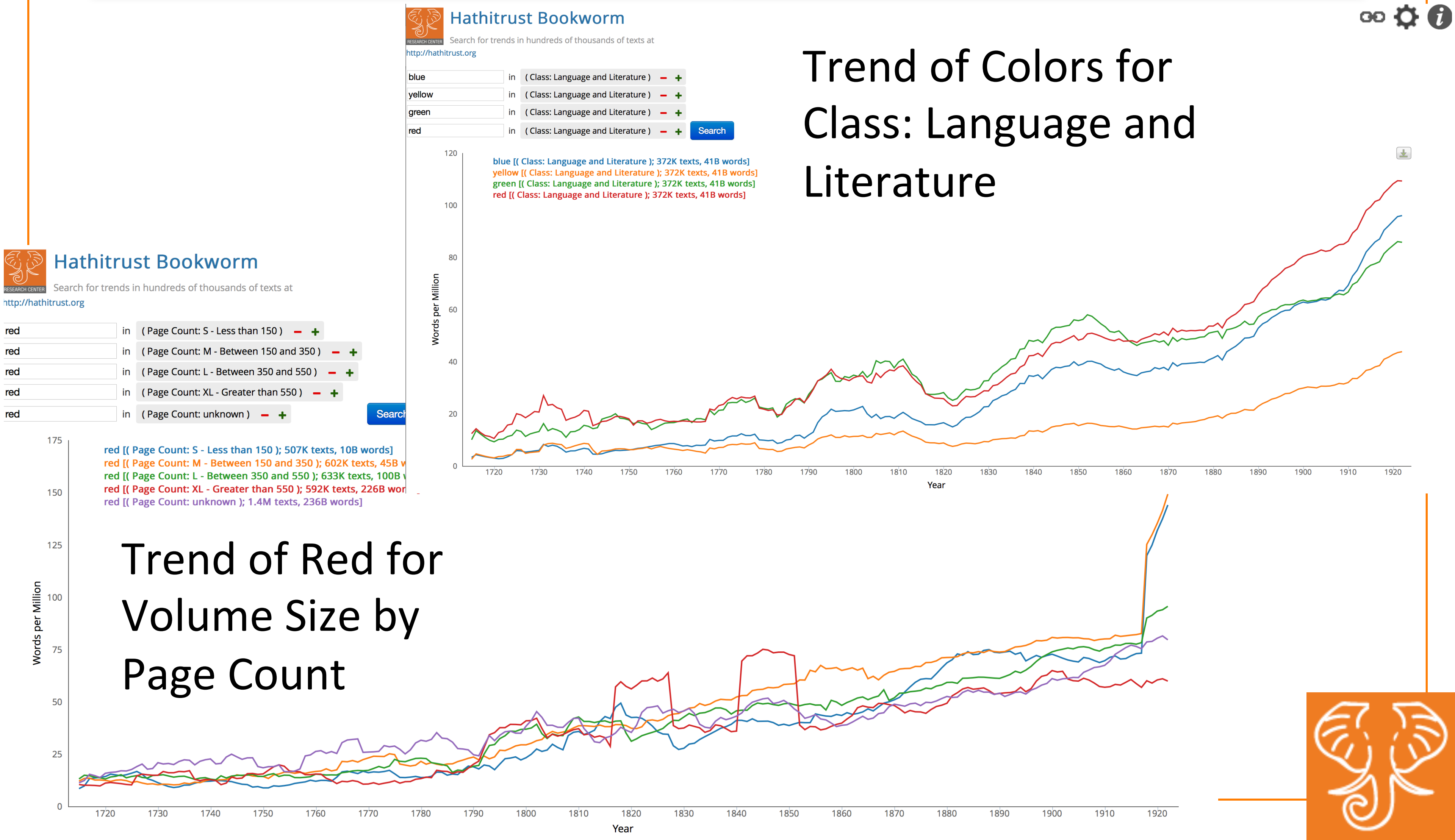
blue [all texts; 3.8M texts, 618B words]

yellow [all texts; 3.8M texts, 618B words]

green [all texts; 3.8M texts, 618B words]

red [all texts; 3.8M texts, 618B words]

Bookworm Adds Facets



Hathitrust Bookworm

Search for trends in hundreds of thousands of texts at http://hathitrust.org

blue in (Class: Language and Literature)
yellow in (Class: Language and Literature)
green in (Class: Language and Literature)
red in (Class: Language and Literature)

blue [(Class: Language and Literature); 372K texts, 41B words]
yellow [(Class: Language and Literature); 372K texts, 41B words]
green [(Class: Language and Literature); 372K texts, 41B words]
red [(Class: Language and Literature); 372K texts, 41B words]

Trend of Colors for Class: Language and Literature

Hathitrust Bookworm

Search for trends in hundreds of thousands of texts at http://hathitrust.org

red in (Page Count: S - Less than 150)
red in (Page Count: M - Between 150 and 350)
red in (Page Count: L - Between 350 and 550)
red in (Page Count: XL - Greater than 550)
red in (Page Count: unknown)

red [(Page Count: S - Less than 150); 507K texts, 10B words]
red [(Page Count: M - Between 150 and 350); 602K texts, 45B w
red [(Page Count: L - Between 350 and 550); 633K texts, 100B w
red [(Page Count: XL - Greater than 550); 592K texts, 226B wor
red [(Page Count: unknown); 1.4M texts, 236B words]

Trend of Red for Volume Size by Page Count

Principal Goals for the HT+BW Project

1. To integrate Bookworm into HTRC in ways that are beneficial to our core demographic of humanities researchers, and
2. To develop our improvements to Bookworm in ways that can be contributed back to the open source project and benefit other large-scale textual repositories.

Tasks

- Implement analytics at scale
  - Development of API for data access
  - Enable SOLR backend in addition to current MySQL
- Identify valuable metadata formats for humanities scholars
  - Development of API for data access
  - Expand metadata available
- Allow creation of custom research collections (HTRC Worksets)
  - Display of trends of only HTRC Workset
  - Create an HTRC Workset from trend viewing
- Generalize beyond HTRC back to Bookworm for usage by others
  - Improvements to GUI
  - API Improvements
- Conduct outreach, training and workshops

Current Metadata

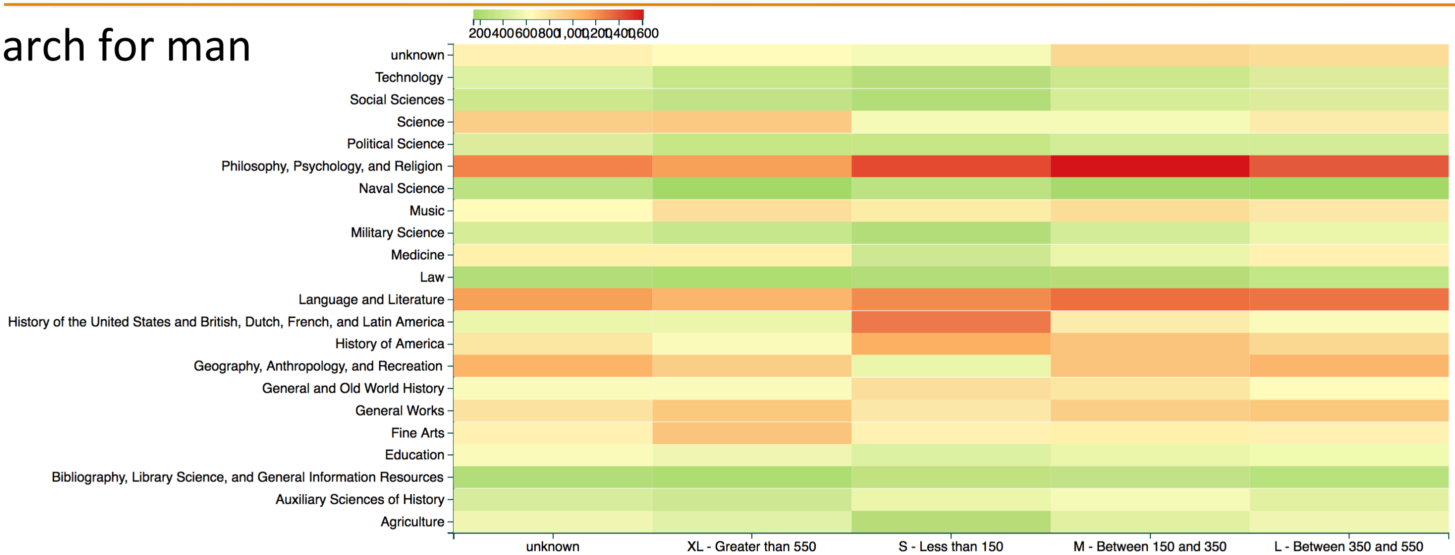
- Class
- Subclass
- Fiction/NonFiction
- Genre
- Language
- Format
- Page Count
- Word Count
- Publication Country
- Publication State
- Publication Place

What additional metadata should we add?

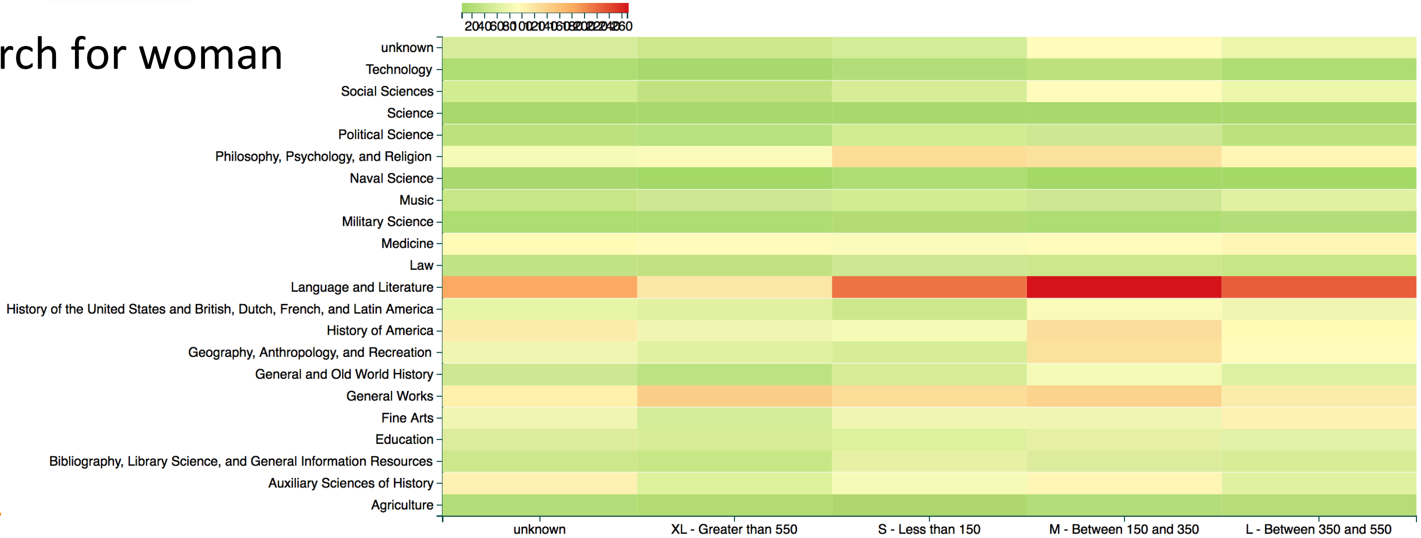
Need hierarchy abilities to make searching more meaningful

API means Data via Heatmaps

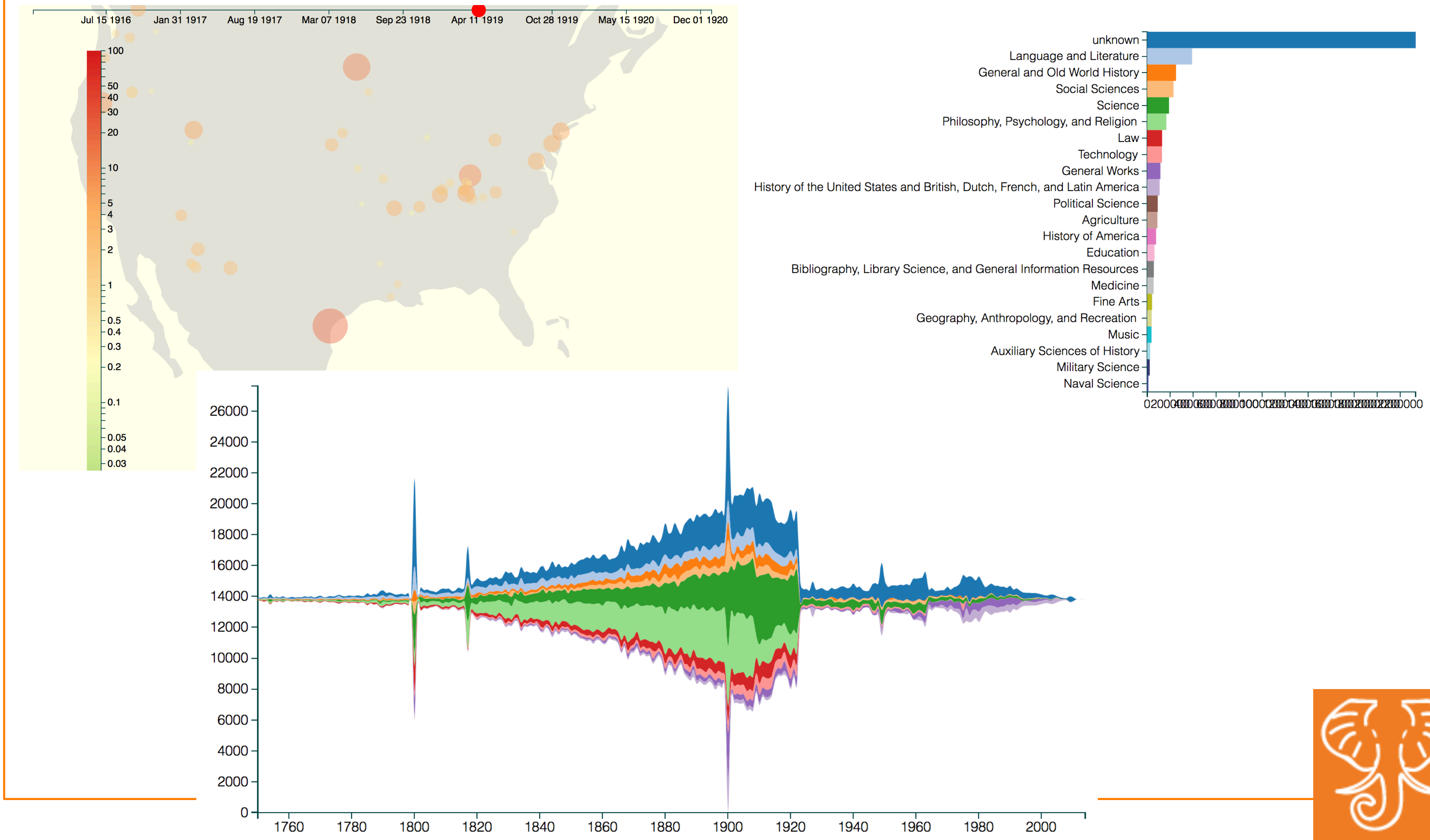
Search for man



Search for woman



API means Other Visual Metaphors



Try it out!

HTRC Bookworm (4.6M public domain volumes)

http://bookworm.htrc.illinois.edu

Other Bookworms

http://bookworm.culturomics.org

Source Code:

http://github.com/Bookworm-project