HathiTrust Research Center - Basic Orientation

Note that hands-on use of the HTRC portal and its tools requires a logon. Please see the information linked from the section titled "The portal", below. Those wishing to experience the tools using a collection of scholarly interest may want to construct such a collection following the tutorial referred to under the section called "Workset builder".

What is HathiTrust (HT)?

- a consortium - international partnership of over 130 institutions.
- a digital library containing about 17 million books, ~6 million (37%) of which are viewable in full online. All items are fully indexed, allowing for full text search within all volumes. You can login with your Cornell NetID to
  - create collections (public or private)
  - download PDF's of any item available in full text
- a trustworthy preservation repository providing long-term stewardship, redundant robust backup, continuous monitoring, and persistent identifiers for all content.

Why aren't <u>all</u> books viewable online?

Computational analysis must address the very real challenges of what can and cannot be legally shared digitally, so it helps to understand the realities that affect full-text viewability. Not all books in HathiTrust are viewable in full, although all are indexed in full. Viewability is determined by many factors, including copyright law (both US and International) and stipulations of the rights-holders (authors and/or publishers) and digitizing agents (like Google). There are two attributes assigned that affect viewability. The first is an attribute that describes a complex set of factors relating to copyright, digitizing agents and rights-holders, referred to as "rights" metadata. The second attribute is a binary value ("allow/deny") often referred to as "access" metadata. In cases where a volume has no factors attached to it that would limit sharing, both attributes would express this. Colloquially, the set of these volumes are referred to as the "open-open" set. What a researcher can do with text is governed by these factors, and the most unrestricted uses can be made from the open-open set.

What is the HathiTrust Research Center (HTRC)?

- a collaborative research center (jointly managed by Indiana University and the University of Illinois) dedicated to developing cutting-edge software tools and cyberinfrastructure that enable advanced computational access to large amounts of digital text. Let's unpack this:
  - "computational access" - computational analysis, algorithmic analysis, distant reading, text-mining
  - "cyberinfrastructure" - for the most part, the Data to Insight Center at University of Indiana: supercomputers, data warehouse, SOLR indexing
  - "large amounts" - "at scale", the bigger the better (better signal, less noise)
  - "cutting edge" - experimental by nature, things can break, things are unfinished/in-development
- intended to serve and build community for scholars interested in text analysis; join user group mailing list (send an email to htrc-usergroup-l-subscribe@list.indiana.edu)

What specific services does the HTRC offer scholars?

Documentation of offerings on the HTRC User Community Wiki - links to services, user support documentation, meeting notes, elist addresses and sign-up information, and FAQs.

The "portal"

- "SHARC" may sometimes be noted: Secure HathiTrust Analytical Research Commons
- access to tools depends on login; see the HTRC Analytics step-by-step tutorial, "Sign up for an account, and sign in" for details

Workset management

- allows researchers to create a set of text to analyze algorithmically, see the tutorial.
- you can create a workset from a file - specification for file are given
- It is a good idea to validate your workset before loading - the validator will let you know if there are issues with your file
- worksets can be private (open to your own use and management) or public (viewable by all logged-in HTRC users, management restricted to owner)

Algorithms

- allow researchers to run ready to use algorithms against specific collections, see the tutorial.
- many algorithms provided (see the full list and descriptions of each) others can be added by scholars' request as time permits development
- workshop dedicated to these alone (ask and I can give you a tour)
- handout available

Data Capsule

- allows researchers to create a virtual machine environment, configure with tools, and analyze texts, Documentation available.
- requires a VNC application for your browser, like VNC View for Google Chrome
- designed to be a secure analytical environment that respects access restrictions to text while allowing for computational analysis; maintenance mode / secure mode
- not yet tied to worksets, but there is a workaround
- currently restricted to "open-open" (non-restricted) corpus; eventual objective is to allow for access to full HT corpus

Datasets

- All data sets including backversions - non-beta offerings
- Extracted Features Data Set
    - A brief single page attachment describing the motivations and potential of the data set.
    - page level attributes (volume level and page level data) for all books in HT; rationale and features explained
    - can download full dataset via rsync (Watch out! BIG! 4TB!)
    - details on leveraging the dataset to select data using a workset and the EF_Rsync_Script_Generator algorithm to download data for just that set.
    - David Mimno's "word similarity tool" is built from the full Extracted Feature data set

Bookworm

- open source project, same basic functionality as Google nGram Viewer, although graphically faceted
- base data is currently linked to a back version of the EF data set that includes 13.5 M volumes both full view and in copyright
- plans and allocated grant to develop tie-in to worksets
- see wiki for tutorial