Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Module 2.1 Gathering Textual Data: Finding Text Lesson Plan

Further reading: go.illinois.edu/ddrf-resources

This lesson introduces the options available to researchers for finding and accessing textual data. In addition to discussing the variety of textual data providers, this lesson covers the process of building a text corpora in the HTDL interface and uploading it to HTRC Analytics for analysis.

Estimated time

35-50 minutes

Workshop audience

Librarians with little-to-no experience with text analysis who may be supporting research and teaching with text analysis at their institutions.

Prerequisites for participants

- Have some idea of text analysis concepts
- Have been introduced to the HTRC, or have completed Module 1

Learning objectives

At the end of the module, participants will be able to:

- Differentiate the various ways textual data can be gathered in order to make recommendations for researchers.
- Evaluate textual data providers based on research needs in order to provide reference to researchers.

- Curate and select volumes to construct their own HTRC workset in order to gain experience building corpora.

Skills

- Build a collection in the HTDL and import it into HTRC Analytics as a workset

Getting ready

- Workshop participants will need:
- An account for HTRC Analytics (https://analytics.hathitrust.org). Instructors may guide participants in the registration process before officially starting the workshop session.

Session outline

- Introduction and outline
- Methods for accessing and downloading textual data
    - Challenges in finding text
    - Sources of textual data
- **Activity:** Strengths and weaknesses of different sources of textual data
- Evaluating sources of text data
- The process of building corpora
- Introduction to worksets
- **Activity:** Create an HT collection and upload a workset to HTRC Analytics
- Creativity Boom case study: How Sam built his corpora for analysis
- Discussion: What expertise do librarians already have to help with building a corpus for textual analysis?

Key concepts

- **Text corpus/corpora:** A “corpus” of text can refer to both a digital collection and an individual's research text dataset. Text corpora, the plural form, are bodies of textual data.
- **Workset**: In the HTRC environment, a workset is a sub-collection of HathiTrust content created by users.
- **Volume:** Generally, a digitized book, periodical, or government document.
- **Optical character recognition (OCR)**: Mechanical or electronic conversion of images of text into machine-readable text. The quality of the results of OCR can vary

greatly, and raw, uncorrected OCR is referred to as "dirty" because it often contains mistakes, while corrected OCR is referred to as “clean”.

Key tools

- **HT Collection Builder**: An interface for creating collections via the HathiTrust Digital Library.

Key points

Kludging access: finding and gathering text	• Text can be approached as data and analyzed by corpus/corpora. • Before analyzing textual data, it is important to ensure the text is of sufficient quality (e.g., OCR-ed data is cleaned up) and fully prepared (certain unnecessary elements are discarded).
Methods for accessing and downloading textual data	• Finding text suitable for computational analysis is challenging, especially with issues of copyright and licensing restrictions, format limitations, and hard-to-navigate systems. • Three commonly used sources to find textual data are vendor databases, digital collections, and social media. Each source has its own strengths and challenges when it comes to downloading text.
Activity: Assess different textual data sources	• In small groups, discuss strengths and weaknesses of different sources of textual data. • *Goal:* Practice assessing benefits and drawbacks of various sources of textual data.
Evaluating textual data sources	• When assisting researchers in finding textual data, also consider how much flexibility is needed for working with the data, the technical skillset of the researcher, and any funding limitations.

Introduction to worksets	• HTRC Worksets are one way the HathiTrust allows users to create text corpora to analyze. • A workset is a user-created collection of HTDL text and can be cited and shared. Viewed on HTRC Analytics, you'll get metadata about the volumes in the workset but will not be able to read the text in this interface, so it suits non-consumptive research. • Users can import worksets from the HT Collection Builder, or compile volume IDs elsewhere.
Activity: Create and import a workset into HTRC Analytics	• Participants work alone or in pairs to create worksets • Encourage attendees to curate the volumes they select for their collection • Whole-group discussion of process when finished • *Goal:* Gain experience using a particular digital library interface to build a text analysis corpus.
Creativity Boom case study	• Introduce how Sam built his corpora for analysis
Discussion	• What expertise do librarians already have to help with building a corpus for textual analysis? • *Goals:* Encourage learners to tie their existing professional knowledge to skills that are useful for building textual datasets.

Additional Tips for Instructors

- **Recommend participants NOT to use Internet Explorer for the web-based activities and choose an alternative browser such as Chrome or Firefox.** Participants using IE may encounter some issues with some of the activities.

- **Make sure to log in to HTDL before creating a collection in HT, and to log in to HTRC Analytics before uploading the collection.**
- When demonstrating activities in web browsers, instructors may use “Ctrl” and “+” (“Command” and “+” on Macs) to enlarge the content on the screen. It can be quite difficult to see things from the back of the room! Use “Ctrl” and “-” (“Command” and “-” on Macs) to zoom back out when you need to demonstrate other things in regular size.