

Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources



Module 5 Visualizing Textual Data: An Introduction Lesson Plan

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

This lesson is an introduction to data visualization in general, with a focus on textual data analysis. It also introduces the HathiTrust+Bookworm interface that allows the user to visualize word usage over time.

Estimated time

30-45 minutes

Workshop audience

- Beginners with an interest in text analytics and/or the HTRC more generally
- Anyone interested in data visualization, especially the visualization of textual data
- Anyone interested in learning about basic tools for interacting with the HTDL corpus

Learning goals

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

- Recognize common types of data visualizations in order to communicate with researchers about their options.
- Explore results in HathiTrust+Bookworm and begin making connections using available data and data points in order to develop experience reading data visualizations.

Skills

- Using library metadata to impact how a visualization is displayed
- Reading and interpreting graphs
- Perform a keyword search
- Fine-tune search results through faceting

Prerequisites for participants

None! While Module 1, Getting Started, provides useful background about the HTRC and its mission, learners can dive into HathiTrust+Bookworm without much introduction.

Session outline

- What is data visualization and when is it used in the research process?
- Common types of textual data visualizations
- **Activity:** Match type of use to the type of visualization
- Examples of web-based tools and programming libraries for visualizing textual data
- Introduction to HathiTrust+Bookworm:
 - What is HathiTrust+Bookworm?
 - Examples of HathiTrust+Bookworm visualizations
 - Overview of HathiTrust+Bookworm interface
- **Activity:** Hands-on exploration of HathiTrust+Bookworm
- Case study: How Sam visualized his data
- Discussion: Visual literacy and data literacy

Getting ready

Workshop participants will need:

- Access to a computer, the Internet, and a web browser.

Key concepts

- **Data visualization:** The process of converting data sources into a visual representation. It often also refers to the product of this process.
- **Word tree:** A type of visualization that displays the different contexts in which a word or phrase appears in a text, with the contexts arranged in a tree-like structure to reveal recurrent themes and phrases.
- **Node-link diagram:** A type of visualization for displaying networks. It captures entities (such as people, places, and topics) as nodes (also called “vertices”) and relationships as links (also called “edges”), with a circle or dot representing a node, and a line representing a link.
- **Word cloud/tag:** A graphical representation of word frequency, usually presenting words that appear more frequently in the source text larger than those that appear less frequently.
- **N-grams:** A contiguous chain of n items from a sequence of text where n is the number of items. Unigrams refer to one item chains, bigrams to two item chains, and so on.

- **Timeline:** A graphic design displaying events in chronological order.

Key tools

- **HathiTrust + Bookworm:** A tool that visualizes word frequencies over time in the HathiTrust Digital Library. It can be accessed at: <https://bookworm.htrc.illinois.edu/develop> .
- **Google Books Ngram Viewer:** Similar to HathiTrust+Bookworm, a tool that enables users to search for words in corpora of texts and visualize their usage over time. Link: <https://books.google.com/ngrams>
- **Voyant:** A tool that can create many types of visualizations including word clouds, bubble charts, networks, word trees, etc. It has a user-friendly interface that works great as a learning tool. Link: <http://voyant-tools.org/>
- **Wordle:** A tool for creating word clouds, mostly for exploration and decorative purposes because not much fine-tuning can be done. Link: <http://www.wordle.net>
- **ArcGIS Online/StoryMaps:** A visualization tool that can be used to incorporate GIS information and maps into interactive timelines and stories. Link: <https://storymaps.arcgis.com/en/>
- **Tableau:** A set of software that can be used for data preparation, visualization, and analysis. Among the different versions of Tableau Desktop (geared towards individual usage), Tableau Public is available for free. See more at: <https://public.tableau.com/s/> and <https://www.tableau.com>
- **Gephi:** A free visualization and exploration software that can be used to create graphs and networks. It works especially well for exploratory data analysis. See more at: <https://gephi.org>
- **NodeXL:** An add-in for Microsoft Excel that supports social network and content analysis. Available in Basic and Pro versions. See more at: <http://www.smrfoundation.org/nodexl/>
- **DH Press:** A digital humanities toolkit that enables users to mashup and visualize a variety of digitized humanities-related material, including historical maps, images, manuscripts, and multimedia content. It can be used to create a range of digital projects and is designed for non-technical users. See more at: <http://dhpress.org>
- **ggplot:** Python library for data visualization.
- **pyplot:** Visualization function in the Python data science package, Pandas.
- **ggplot2:** R library for data visualization.
- **D3.js:** JavaScript library for web-publishable visualizations.

Key points

What is data visualization?	<ul style="list-style-type: none"> • Data visualization is the process of converting data sources into a visual representation. • Visualization is a way of interpreting and presenting data.
Common textual data visualizations	<ul style="list-style-type: none"> • Some common visualizations include: word clouds, trees/hierarchies, networks, temporal/spatial-based visualizations, and other “multi-dimensional” visualizations.
Activity: Match type of use to the type of visualization	<ul style="list-style-type: none"> • Participants match types of visualizations to the kinds of information they are suited to convey. If time allows, consider the kind of data each visualization might require. • <i>Goal:</i> Practice thinking about applications for data visualization, and when and with what data they might be employed by researchers.
Examples of web-based tools and programming libraries for visualizing textual data	<ul style="list-style-type: none"> • Examples of web-based tools include: Voyant, Wordle, ArcGIS Online/StoryMaps, Google Books Ngram Viewer, HathiTrust+Bookworm, Tableau, Gephi, NodeXL, DH Press • Programming libraries for visualizations: matplotlib, pyplot, and ggplot library in Python; ggplot2 in R; D3.js.
What is HathiTrust+Bookworm?	<ul style="list-style-type: none"> • Bookworm is a tool that visualizes language usage trends in repositories of digitized texts. It is good at finding and understanding categories in a library. • Bookworm can visualize and quantify the dynamics of language evolution. • HathiTrust + Bookworm is a visualization of word frequencies over time in the HathiTrust Digital Library.

Examples of HathiTrust+Bookworm visualizations	<ul style="list-style-type: none"> Using HT+BW to track social change: “lady” vs. “woman” Using HT+BW to Bookworm to track words in translation across time and place: “liberté” and “liberty”
Overview of HathiTrust+Bookworm interface	<ul style="list-style-type: none"> Type in search words and click on the funnel icon to facet the search by genre, language, and more. Use the tabs “Dates”, “Metric”, and “Case” to fine-tune results. After the visualization is generated, click on a specific spot on the curve to be directed to corresponding volumes in the HathiTrust Digital Library.
Activity: Hands-on exploration of HathiTrust+Bookworm	<ul style="list-style-type: none"> Guide participants in using HT+BW to visualize lexical trends. <i>Goal:</i> Gain experience using web-based visualization tools, the parameters that can be adjusted, and the information they convey.
Case Study	<ul style="list-style-type: none"> Sam used HT+Bookworm to visualize the use of “creative” in the HTDL over time Sam also used an experimental HT+BW interface to create different kinds of visualizations
Discussion	<ul style="list-style-type: none"> Where does visual literacy fit into data literacy overall? What would it mean to be visually literate, particularly with regard to text analysis? <i>Goal:</i> Encourage librarians to consider pedagogical applications for concepts they have learned.

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.

- 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.
- For the HT+BW hands-on activity, instructors may encourage workshop participants to discuss their search results with each other. This can make the activity more interactive and keep the participants more fully engaged.
- Data visualization is a huge topic, and the information provided in this lesson can only scratch the surface. For instructors who have little previous experience in this area, it may be helpful to do some additional background reading (the materials provided in the further reading section of our website is a good place to start) to familiarize themselves with other types and formats of data visualization and more visualization tools.