Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Module 2.1 Gathering Textual Data: Finding Text Instructor Guide

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

Narrative

The following is a suggested narrative to accompany the slides. Please feel free to modify content as needed when delivering workshops. Additional information and examples that are provided to aid the instructor's understanding of the materials are in *Italic text* – whether to present them to workshop audiences is entirely up to the instructor.

Slide M2.1-1

This lesson introduces the options available to researchers for 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 importing it into HTRC Analytics for analysis.

Slide M2.1-2

Here is an outline of what we will be covering in this module.

- First, we will explore options for finding data for text analysis.
- We will also consider the concept of a dataset for text analysis.
- For our hands-on activity, we will build our own corpora of HathiTrust text and import it into HTRC Analytics for analysis. Note here that when viewing text as data, we usually analyze them by corpus or corpora. A "corpus" of text can refer to both a digital collection and an individual's research text dataset. Text corpora are bodies of text.
- Finally, we will also take a look at our case study and learn how Sam built his Creativity Corpus of HathiTrust volumes.

Slide M2.1-3

By the end of this module, you will have created a collection of volumes from the HathiTrust Digital Library and prepared it for analysis in HTRC Analytics as a workset.

Slide M2.1-4

Thomas Padilla, Visiting Digital Research Services Librarian at the University of Nevada Las Vegas, wrote about three common challenges among text analysis projects.

- **First**, data of interest must be found. **Second**, data must be gettable. **Third**, if it's not already formed according to wildest dreams, ways must be known of getting data into a state that they are readily usable with desired methods and tools.
- Before any text analysis can be conducted, researchers need to find the textual data that they need, and this process is not as easy as it may seem. In this module, we will be exploring different options of finding text.

Slide M2.1-5

There are plenty of sources for finding textual data for analysis, but suitable text for computational text analysis is not always easy to get.

- One issue that a researcher may run into is copyright and licensing restrictions, which may prohibit users from storing or publishing data. *This is especially common when you are looking for texts provided by vendor databases.*
- Another common issue is texts are often provided in formats that require additional processing before computational analysis can be conducted. For example, some documents are only provided as scanned images, which will need to be OCR-ed and cleaned before analysis.
- A third issue is many systems that a researcher may want to obtain text from were not built with text analysis in mind, so they can be difficult to navigate and/or challenging to pull text from. Higher level programming and tech skills will be needed in such cases to get the text out.
- Additionally, such issues can be easier to address or circumvent when you are only working with a small amount of text, but when researchers want to work on a larger scale, the difficulties get compounded. Hopefully, things will get easier in the near future.

Slide M2.1-6

Now, let's look at three common types of sources where users can start looking for the textual data that they want.

- One way to get data is from vendor databases provided by libraries. Data provided by vendors can be of higher quality, but this is also where one needs to be very careful about licensing restrictions. Often, a library's agreements with vendors may prohibit the steps needed to do text analysis, such as storing or publishing data.

There are a few strategies that can make getting text with vendors easier:

- A library can modify or add to their contract with the vendor to get access that permits text analysis. *This may take some time and additional negotiations on the library side, but the pay-off can be very valuable to users.*
- Some vendors are currently building tools and services for text mining their collections, so obtaining text may be easier in the foreseeable future. *However, it is unclear how long it will take for a specific vender to get these tools and services fully developed, and not all vendors are doing this.*
- In some cases, researchers can ask for special permission to mine a specific portion of vender-provided data, but sometimes the vendor will require fees for doing so.

One example of existing vendor-provided services is JSTOR Data for Research (http://dfr.jstor.org). *This is a free service for researchers wishing to analyze journal and pamphlet content on JSTOR, and the website provides data sets of OCR, metadata, key terms, N-grams and reference text of a portion of JSTOR content. By creating an account, users can download metadata, word frequencies, citations, key terms, and N-grams of up to 1,000 documents. Larger-scale requests are handled on a case-by-case basis.*

Slide M2.1-7

Digital collections (from libraries, archives, and museums) often have a wealth of text data. However, data from these sources are usually siloed, and access is not formulated for research at scale.

When finding data in digital collections, there are two things to look for that can make things easier. First, try to get text that is in an accessible format, such as plain text. Second, look for interfaces that can permit bulk download.

UNC's DocSouth Data is an example of a digital collection that can be used for text analysis. There are many others.

DocSouth Data (http://docsouth.unc.edu/docsouthdata/) provides access to some of the Documenting the American South collections at the University of North Carolina in formats that

work well with common text mining and data analysis tools. Currently, four collections are available for download in a zip file containing each of the texts in the collection as plain text, each of the texts with complete TEI/XML markup, a Table of Contents file, and other supporting documents.

Slide M2.1-8

Social media platforms are an additional source of text data, and research based on analysis of social media data is getting increasingly popular and useful.

Currently, some social media platforms have designated systems that allow access to some data. There are also some 3rd-party tools available. Twitter in particular has an API.

An API (Application Programming Interface) is a set of clearly-defined communication methods (may include commands, functions, protocols, objects, etc.) that can be used to interact with an external system. They are basically instructions (written in code) for accessing systems or collections. We'll come back to APIs later.

Slide M2.1-9

Let's do a short activity before we move on to the next section. We just introduced three different sources of textual data and some of their characteristics. We would like you to consider in pairs or in small groups: If we are building a corpus for political history, what are the strengths and weaknesses of each of these broad sources for textual data? Please discuss and take some notes in the chart together (it's provided in the handout of this section), and we'll come back together as a group and share our ideas.

(Instructor organizes the activity. If time runs short, only discuss in pairs and skip the group sharing.)

Slide M2.1-10

When assisting researchers in finding textual data for their projects, there are several things to consider about evaluating a textual data source.

Apart from clarifying the period/place/or people of interest and if the researcher already has a data source they'd like to use, you might also want to assess:

- Does the researcher already have a data source they'd like to use, and is it digitized?

- Are there potential copyright and licensing restrictions? Does the researcher study the mid-20th century? Do they want to use a journal database? If so, look for services that try to help researchers legally circumvent these restrictions.
- Do they have a period, place, or person of interest - that will obviously heavily dictate what data source they will find useful?
- You may want to know more about how much flexibility is needed for the researcher working with the data. If the researcher wants to do a lot of their own manipulation and clean-up, then you should look for sources that are as open as possible, like the Internet Archive.
- It is important to consider how technically experienced is the researcher. If the person is a beginner, you probably want to let then try looking at sources with easy bulk downloading, like DocSouth Data.
- It will be helpful to know if the researcher has funding. Sometimes you have to pay extra for access to OCR text from vendor sources, so it is better to get an idea if the researcher has any resources that can be allocated to paying for textual data.

Slide M2.1-11

Next, we will talk about the process of building corpora. As mentioned before, text corpora are bodies of text.

- A researcher starts building corpora by first identifying texts that fit their specific research needs. This usually involves finding texts that contain a key term or phrase through a full text search.
- According to the specific research question, one can also identify texts through metadata, for example searching for texts written by certain author(s), that are within a particular date range, or of a specific genre.
- The identification process can, of course, also be some combination of full-text and metadata search.

Slide M2.1-12

- The process will also typically involve doing some deduplication, which means getting rid of multiples of volumes.
- What to keep and discard will be dependent on the specific needs of the project.

- Some examples of deduplication a researcher might choose include: choosing the volume with the best OCR quality, choosing the earliest edition, or choosing editions without forewords or afterwords.
- *We will be discussing a little more about this when we get to Sam's project later in this module.*

Slide M2.1-13

Now, let's look specifically at how the HTRC can provide access to textual data and assist users in building up a text corpora for analysis.

- HTRC Worksets are one way the HathiTrust allows users to create text corpora to analyze. A workset is a user-created collection of texts from the HathiTrust Digital Library. You can think of them as textual datasets.
- The idea behind worksets is that researchers can build their own collections, which is something they are used to doing also in the analog world, and these worksets can be cited by and shared with other users, which helps encourage ensuring reproducibility (for example, other people can test your results when they run the same analysis on your shared worksets).
- Worksets are also suited for non-consumptive access, which you may remember means users cannot directly "consume" the text in full, but can run tools or algorithms against data for research analysis.

Slide M2.1-14

When you view a workset on HTRC Analytics, you'll get metadata about the volumes in the workset. You cannot read the text in this interface. At its core, the workset is just a manifest of volume IDs.

Slide M2.1-15

How can we build worksets? First, worksets are stored in HTRC Analytics, and you will need to create an account linked to an institutional email address to store your worksets.

Currently there are two ways to build a workset:

- You can create collections with the HT Collection Builder and import them into HTRC Analytics as worksets. We will be building a workset using this method in our later hands-on activity.
- Another way is to compile volume IDs elsewhere and upload it to HTRC Analytics, and HTRC Analytics will create a workset based on that list.

Slide M2.1-16

Let's go back to our sample reference question. The student would like to examine how concepts such as liberty have changed over time, so one way we can approach this question is to first create a textual dataset of volumes related to political speech in America with the HT Collection Builder, and upload it to HTRC Analytics as a workset for analysis.
The student has referenced a series of publications called "Public Papers of the Presidents of the United States" previously in their research. They think these volumes, which contain presidential speeches, would be a good data source for their research.

Slide M2.1-17

For our hands-on activity, let's practice building worksets. We will log in to HTDL and create a collection containing volumes of the public papers of the presidents of the United States, and then import it into HTRC Analytics as a workset. Use the links on the slide to access HTDL (https://www.hathitrust.org) and HTRC Analytics (https://analytics.hathitrust.org). **Make sure you are logged in to HTDL before creating a collection.**

(Instructor may choose to demo live/go through the screenshots for the first few slides of building a collection, then let participants work on their own to create the collection. Afterwards, the instructor can demo live/go through the import process in HTRC Analytics along with the participants.)

Slide M2.1-18

(Slides included in case of technical issues, opt for live demo when possible.)

First, let's go to the HTDL interface: https://www.hathitrust.org

Slides M2.1-19 to M2.1-20

Click on the "LOG IN" button on the right to sign in. If you are affiliated to an HT partner institution, select your partner institution from the list and click on continue, then follow the directions for institutional log in. If not, click on "See options to log in as a guest", and you will be directed to a page with multiple options (such as logging in with your Google or Twitter account). **Make sure you are logged in to HTDL before creating a collection.**

Slide M2.1-21

Once you are logged in, click on the "FULL-TEXT" tab to search in full text. Now you can start thinking about a search query to find volumes for your collection. Why don't you try building your

own collection for our imagined researcher? You can check the instructions on the handout if you get stuck. We'll give you 5-10 minutes to make your collection, and then we will regroup.

(Optional narrative below)

- *First, enter in a search query.*
- *For this example, let's do an advanced full-text search for volumes that contain both "public papers" and "United States" in their titles. Because I'm an expert librarian, I know there are published speeches by presidents called, "Public Papers of the Presidents of the United States" so I'm structuring my query to look for those volumes.*
- *Note that the default setting is doing a simple search in full-text, so you will need to click on the "Advanced full-text search" link under the search bar.*

Slide M2.1-22

(Optional narrative below)

You will see two blank search fields connected by "AND" logic for you to fill in. Since we want to look for volumes that have two specified phrases ("public papers" and "United States") in their titles, the "AND" logic works perfectly and we don't need to change it. For both search fields, select "this exact phrase" and "Title" to limit our search. Type "United States" in one search field and "public papers" in the other (without the quotation marks). Click on "Search" at the bottom of the page to view results.

Slide M2.1-23

(Optional narrative below)

After the search results appear, we can further facet the results using the options provided in the sidebar on the left. After you are satisfied with the results, click on the "Select" box on the left side of an entry to select it for your collection. You can also click on "Select all on page" when you feel like every item on the page should be added to your collection. When you're ready, click on the "Select Collection" bar and choose "[CREATE NEW COLLECTION]" from the drop-down menu. Then hit the "Add Selected" button on the right.

(For instructors, we recommend that you only select a couple of items using the check boxes during your demonstration, since it may take a long time for the system to process your request if you select all items or many items. You may want to recommend workshop participants to refrain from selecting too many items as well during the activity so no one will need wait too long

to get to the next step. Encourage attendees to take a curatorial perspective when selecting volumes.)

Slide M2.1-24

(Optional narrative below)

When you're satisfied with your selection, click on the "Select Collection" bar and choose "[CREATE NEW COLLECTION]" from the drop-down menu. Then hit the "Add Selected" button on the right.

Slide M2.1-25

(Optional narrative below)

A pop-up window will appear that prompts you to add some metadata to your collection before the system creates it for you.

Fill in the name and description of your new collection. You can choose to make the collection public or private, and we recommend writing a short description whenever you make collections public. When done, click on "Save Changes" to create your collection.

Slides M2.1-26 to M2.1-27

(Optional narrative below)

After the collection is successfully created, you should see a confirmation message above the search results. To view your collection, click on "My Collections" near the top right of the page.

This will bring you to all your collections. You can manage your collections here by viewing collections, changing the public or private status of a collection, and deleting collections you don't need. Click on the title of the collection you just created to view it.

Slide M2.1-28

(Optional narrative below)

You will be able to see the title and description of your collection, as well as all the items in it. Copy (highlight and ctrl-c/cmd-c) the URL in the bar. You will need this to create your workset.

Slide M2.1-29

Now, we will need to go to HTRC Analytics and import our collection as a workset for analysis. This is what the HTRC Analytics website looks like. We suggest logging in first after landing on this page.

Slide M2.1-30

Look for the sign in/ sign up section on the top right part of the page. Since we've already created an account, let's click on "sign in".

Slides M2.1-31 to M2.1-32

We are now brought to the login page. Enter your user name and password, then click on the "Sign in" button.

After you've logged in successfully, you will be directed to the homepage again and your username will appear on the top right.

Slide M2.1-33

Now, we are ready to import our HT Collection into HTRC Analytics as a workset. Click on the "Worksets" option on the header menu. This will bring you to the worksets page.

Slide M2.1-34

You can view your own worksets as well as public worksets on this page. To import your HT collection, click on "Create a workset" near the top right.

Slide M2.1-35

There are two creation options for HTRC worksets: either upload a file containing a list of HathiTrust volume IDs if curated outside of the HTDL interface, or import from the HTDL using the collection URL. This activity uses the "import from HT" method.

Slide M2.1-36

Paste (ctrl-v/cmd-v) your HT collection URL in the Collection URL field and click to retrieve the information for the collection. Name your workset and write a description. When naming, please note that only characters A-Z, 0-9, (), -, or _ are allowed, so do not use spaces or other special characters. Check the "Make private workset" box if you want to create a private workset. Finally, hit the "Create Workset" button.

Slide M2.1-37

If successful, you should be brought back to your worksets page, and the new workset that you just created should be listed. Click on the name of the workset to view it.

Slide M2.1-38

Let's wrap-up our activity and talk about what you did.

- How did it go?
- What kind of search criteria did you use?
- Did you find any challenges?

Slide M2.1-39

Finally, let's look at our case study and see how Sam completed the step of finding the textual data that he needed for his project.

- With our help, Sam did a search across the HTDL for the term "creativ*", which would give him results that contained either "creative" or "creativity".
- After the search, he had an initial list of 2.7 million volumes, but there were some duplicative materials.
- For the purposes of his project, he decided that duplicates may impact and skew his results, so he moved on to the deduplication process.
- In the end, he decided to keep different editions of same work but discard multiple copies of same edition. It should be noted that this worked for his project, but different projects may require different rules when doing deduplication.
- Deduplication was not done "by hand" but was done using an automated process to compare metadata.
- Sam ended up with a final, refined list of volumes (a workset) which was his "creativity corpus."

The decision to discard identical volumes or editions, while preserving multiple editions of a single title, was made with the idea that more widely published works are also more widely circulated and therefore more influential and/or paradigmatic of the way language is used in general.

Slide M2.1-40

Let's wrap up this module with a discussion. What expertise do you think librarians already have to help with building a corpus for textual analysis?

(Short discussion for 5-10 minutes. if time runs short, skip this discussion.)

Slide M2.1-41

That's all we have for this lesson, and we will be happy to take any questions from you.

Slide M2.1-42

(*Display references so attendees know they can find them later.)*