Humanities and Research Data

Caterina Agostini

This workshop is designed to introduce some ways of working with data in the humanities by transforming sources into machine-readable formats. We will learn ways to organize data and explore some of the main digital methods.

While we often handle research materials, digital and not, translating those sources into usable data formats is a process where scholars decide what aspect(s) to study, therefore emphasizing certain perspectives. We will see some elements to consider in a digital project, learn from case studies, and discuss some of the benefits (and costs) of translating materials into machine-readable data.

What Types of Materials?

For this workshop, I have selected some sample materials in the humanities. To start, we possibly have a variety of materials to transform into data, in particular, textual sources (from libraries and digital collections), and visual sources (libraries and museums).

Digitized Texts

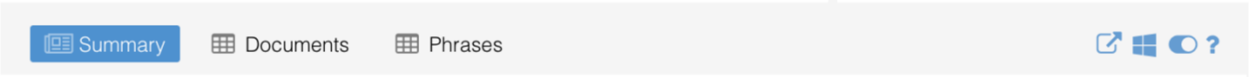
Starting with texts already available in digital forms, we will look up one book on the Internet Archive https://archive.org/, Hathi Trust https://www.hathitrust.org/, or Project Gutenberg https://www.gutenberg.org/

For example, I have searched for one edition of the works by William Shakespeare
https://archive.org/details/theworksofwillia00shakrich
Next, we will download the .txt file.

Exploratory Text Analysis

If you are curious about text analysis, a good starting point is an online resource called Voyant Tools. We will run some exploratory investigations through Voyant Tools here: https://voyant-tools.org/
Copy and paste the text that you have just downloaded, or upload one or more text files from your computer. Once you have a corpus (one or more files) uploaded to Voyant Tools, the default configuration of tools will show the following options: Cirrus, Reader, Trends, Summary, and Contexts. We will look at each of those options.

Next, we might want to select from pre-existing "stopword" lists in various languages or make our own. To access the stopword option, click on the lower left corner (where you see a blue button that looks like an on/off switch). In many cases, you do not need to select a list or a language, as Voyant will auto-detect the language of the text and automatically select a stopword list. If you do not want a stopword list, you can select None. This will produce a dialogue box where you can select from one of the existing stopword lists. All lists are customizable, so you can select a language and then click on the Edit List button to modify words. In any case, clicking on the Edit List button is also a good way to read what words are in any stopword list, regardless of your wanting to use or edit it. Stopword lists are composed of one "word" per line, and there might be room for interpretation of what a word is here.



The above screenshot shows the main buttons you need to see and add tools, and to check stopword lists.

After running your project, you can bookmark and share URLs that refer to your collection of texts. You need to click in the upper right corner, on the button that looks like a box with an upward arrow. This also allows you to work with the same texts during different sessions, without having to reload all the documents each time, and a corpus will remain accessible as long as it is accessed at least once a month. You can export a link for your corpus and the current set of tools clicking on the "Export" (diskette) icon in the blue bar at the top, or export a link for an individual tool by clicking on the "Export" icon in one of the tool panes.

Datasets

Let's see some examples of datasets:

- A numismatic dataset from MANTIS that includes the Roman, Byzantine, and Islamic coins from between 400 and 500 CE. This is an interesting collection because it uses regions and the earliest, or latest possible date for items https://raw.githubusercontent.com/leadr-msu/hst251-NumismaticViz/master/MANTISRomanByz Islamic400-500-cleaned.csv
- A dataset on the French "encyclopédistes" shared by Melanie Conroy https://ageofrevolutions.com/2017/09/25/visualizing-social-networks-palladio-and-theencyclopedistes-pt-i
- A dataset from the Getty Provenance Index Databases https://github.com/chnm/doingdh

Organizing Information into Spreadsheets

Organizing information into a spreadsheet is a recommended step to build a digital project around humanities so that you can have research data. Why do we need a spreadsheet from research materials? One reason could be to visualize what you have found, through a map, network analysis, a graph for trends or value distributions, and to see the implications of your claims around the data. Exploratory questions can be addressed in http://hdlab.stanford.edu/breve/, a tool currently developed for the analysis of tabular data.

A spreadsheet is a file in which we organize information into categories.

- Rows are items of interest (people, objects, court cases, places…)
- Columns are attributes of the items (birth town, cost, date a law was established, status, language spoken, gender…)
- Cells are the boxes

Possible file formats for spreadsheets are .xls and .csv. A comma-separated value file is particularly useful to build digital projects, for example through Palladio, a web tool for mapping: http://hdlab.stanford.edu/palladio/.

To see one example, let's find a list of passengers from the Ellis Island records https://www.statueofliberty.org/discover/passenger-ship-search/

Let's now compare this set of information with other historical sources available online. https://www.archives.gov/nyc/finding-aids/passenger-lists.html#online

From a catalog of disciplines and faculty in a university in the Renaissance and early modern period, we will see how information is organized and presented. What questions arise? https://www.google.com/books/edition/Gymnasium_Patavinum_Jacobi_Philippi_Toma/nFEVAAAAQAAJ?hl=en

Now we have information about passengers or scholars. What categories do we notice? How can we transcribe information?

Recommended

Always keep a separate, original copy of the original, unstructured data for publication, reference, and future uses.

Always create another spreadsheet with your cleaned, structured data for visualization with columns and formats that satisfy the tool(s) you want to use.

Data should be uniform, consistent, and simple. Document all changes you make to the data to achieve this.

Build a spreadsheet

Let's go from unstructured to structured data. Every column represents an attribute that can inform your analysis.

Diacritics, Punctuation, Special Characters, and Non-Latin Script

Avoid diacritics, punctuation, special characters, and non-Latin script if you can
If you can't, be sure to save your files in UTF-8 encoding (Google Sheets does this automatically) to ensure your documents/data can be understood by a large variety of tools
Excel for Mac does not natively support the import or export of UTF-8 encoded files
Import the file into Google Sheets, then save as .csv

Create a separate column with the diacritics, punctuation, special characters, and non-Latin script items if necessary

Synthesizing data columns

To create a new column of combined information, first create the new, blank column
Name it
Use this example function:
=CONCATENATE(A2,”,”,B2)
The combined item from A2 and B2 will have a comma and a space separating them
Split data columns for more granular exploration
For any attribute that groups together sub-attributes, split these into the parts
Select the column to be split
Data > Split text to columns
Excel or Google Sheets will try to understand your data and present it in an accountant friendly way
Format Cells, [example]

Column headers

Each column must have a header and be unique
Do not have a gap between the column header and the data, or any empty columns, and/or rows in your table
Do not use spaces, hyphens, or underscores

Empty or Null Cells

Do not leave cells empty
Have a consistent method for noting No Answer, Null, or Missing values so these cells are not mistaken as zeros or otherwise misinterpreted

If you need to guess or estimate, create a column to write out your rationale, and consider another column with a percentage, or score indicating how certain you are of your guess or estimate:
for example, rough guess, source derived, contextually likely, source likely, mean of range.

Consistent dates

Dates should be entered in the following format: Year-Month-Day (2020-11-18)
Years must always be entered as four digit integers. Thus, the years between 0 and 1000 should be rendered as 0001 and 0999
BCE dates should follow the same format, but preceded by a minus sign (0200-01-01)
Be sure to specify that the column reads as “Text” rather than “dates” (Format Cell > Text), so Excel/Google Sheets does not change the dates into its own format
=TEXT(B2, “yyyy-mm-dd”)

LatLong and decimal degrees

Location information must be rendered as latitude and longitude
Use three geocoded columns to work around the preferences of various mapping tools: LatLong, Latitude, and Longitude
If you have a list of place names, but no associated coordinates, you can use this geocoding tool: https://gis.ucla.edu/geocoder, or Google Maps for individual location information

In Google Sheets, you can also try Geocode by Awesome Table (Add ons > Geocode by Awesome Table > Start Geocoding)

If no modern address is associated with your location, you can manually pin a location on a Google Map to get this information, which is useful for sites with no address.

Images

You should create a column in your spreadsheet with the URL to an image you want to appear in a pop-up. The URL must be publicly accessible and reference the image, not the page with the image.

Transcription Tools

Transkribus.eu is being developed to recognize handwritten records and transcribe them into digital form. It requires 100 pages of transcription to create an engine unique to your scribe.

Another tool under development is Data Scribe, an Omeka S plugin:
https://rrchnm.org/news/from-historical-sources-to-datasets-a-preview-of-datascribe/

Recogito (https://recogito.pelagios.org/) can be used to find location information in digital texts. It runs its Name Entity Recognition Engine on your text, using English, but other languages and engines can be integrated, and it correlates words in your text with geographic information from four gazetteers.

To Map and Annotate Your Materials

You can use Recogito for individual or team projects, when you need to collect and categorize people, places, events, or other elements that you can add as annotations. Mapping is an option, too, given that geographic information is available through built-in gazetteers that the Recogito team connected to the web tool.

Preserving and Documenting Digital Work

The Digital Documentation Process (DDP) guides scholars through a catalogue record, a persistent identifier, and an archiving dossier narrative. The DDP guidelines and project were developed by Dr. Laura Morreale, so that following the DDP allows to make DH scholarship findable and citable, durable, and it enhances the value of DH work. Go to https://digitalhumanitiesddp.com/ for more information. I would like to express my gratitude to Dr. Laura Morreale for academic training and mentorship in digital humanities methods and tools, some of which I have shared with you here.

You can also share your publications and presentations on an open-access repository such as https://zenodo.org/

Resources

An online guidebook for Voyant Tools is available here: https://voyant-tools.org/docs/#!/guide/start
More on the Voyant Tools visualizations here: https://voyant-tools.org/docs/#!/guide/tools

A quick how-to guide introducing Recogito is online at https://recogito.pelagios.org/help/tutorial
For more digital project concepts, see the Cambridge Digital Humanities website: https://www.cdh.cam.ac.uk/lab/concept-1

More on digital tools and methods at the Rutgers DH community: https://libguides.rutgers.edu/dh-lab

To learn more on quantitative textual analysis in R, see the introduction by Alex Leslie https://github.com/azleslie/TextAnalysisIntro

For teaching purposes, see a blog post by Caterina Agostini: https://dh.rutgers.edu/digital-humanities-tools-in-online-humanities-classes/

Contact Information

You can contact and follow up with questions:
Caterina Agostini, Rutgers Italian Department, caterina.agostini@rutgers.edu