LESSON PLAN:
Building forms to transcribe structured historical data

The activity should be part of a larger conversation about structured data, historical data, and how humanists create machine-readable datasets.

The activity will help students understand the decisions which are involved in going from a historical source with structured data to a quantitative analysis or visualization. They will learn about the options (like controlled vocabularies) which can help ensure tidy data but might also create omissions in the dataset.

Activity

This activity has students work through the brainstorming part of creating a form for transcribing a historical source using the options provided by DataScribe.

Materials

- Primary sources which have structured data. These can be digital or printed handouts.
- Form building worksheet (attached)
- Colored pencils, highlighters, markers (optional, for use with printed sources)

Preparation

Prepare copies of the primary sources for which students will build forms. Use one source for all students or a series of similar sources.

Possible sources include:

- 1950 census forms <https://1950census.archives.gov/howto/blank-forms.html>
    - Especially those for Alaska, Hawaii, American Samoa, and Guam
- Historical city directories:
    - New York city directories from the New York Public Library <https://digitalcollections.nypl.org/collections/new-york-city-directories#/?tab=about&scroll=12>
    - There are many digitized city directories available through the Internet Archive
- Ledgers, Registers, and sAccount books

    - Mason family account book, C0214, George Mason University. Libraries. Special Collections & Archives <https://scrc.gmu.edu/finding_aids/masonaccountbook.html>
    - Miller, Warum. Ledger. 1770, https://doi.org/10.5479/sil.396039.39088006598395
    - United States. Surgeon-General's Office. Meteorological register for the years 1822, 1823, 1824, & 1825. Washington: Printed by Edward De Krafft, doi: https://doi.org/10.5479/sil.453007.39088007566821
- Associated Press news dispatches <https://www.loc.gov/collections/associated-press-news-dispatches-1915-to-1930/about-this-collection/>. See the *Plague in Iqueque* case study <https://datascribe.tech/resources/casestudies/#plague-in-iqueque> for a model of working with semi-structured data.

Opening Discussion

Start by grounding the creation of the form in the desired outcome: a clean dataset which can be used for computational analysis. In order to create that dataset, the scholars (students) need to consider the following questions:

- What information do I need to capture for the analyses I want to do?
- What outputs will be most helpful for that analysis?

Introduce the students to the specific sources they will be working with. Optional: give them an imagined research project or have them come up with a possible research question which might use the data from this source.

As a class, walk through the data types which can be used when building the form. Be sure to mention the options for data types, for example maximum and minimum year.

Activity

Break the students into groups of between three and six people. Give each group a source and a copy of the blank worksheet (link).

Optional: If the sources are printouts, encourage students to use markers, highlighters, colored pencils to mark up and annotate the primary source printout, identifying what data they want to capture.

Each group should work together to create a form for their source. They should designate at least one field as the primary (identifier) field. Encourage them to use the notes field to explain or document the settings they would use - for example, options for a select or radio button. Also ask them to make notes on why they chose each data block option.

Give each group at least 20 minutes to work on their sheet.

Bring the students back together in one group. Ask each group to report on the process. Questions might include: Did they create one form for the entire source or break it into parts (for more complex documents like the census)? What information in the source did they omit, if any, and why? What controlled terms did they come up with? What do they expect might be confusing to transcribers? What challenges did the source present and how did they consider resolving them? Were they disagreements about how to handle specific data points? What field(s) were required and why? What field did they set as primary and why?

Encourage students to reflect on how the decisions they make when creating the form would shape the resulting datasets, and therefore the analyses resulting from working with those datasets.

Possible assignments

- Have students read a DataScribe case study <https://datascribe.tech/resources/casestudies/>
- Have students explore some structured historic datasets, compare original sources (if they can find them) to what's been made public either through datasets or visualizations

Appendix 1: Examples of humanities projects using structured historical data

American Religious Ecologies (https://religiousecologies.org). American Religious Ecologies seeks to understand how congregations from different religious traditions related to one another by creating new datasets, maps and visualizations for the history of American religion. While some Americans have lived in rich religious ecologies, surrounded by a plethora of denominational choices, others have lived in places with only one or a few religious options. Using new and existing datasets, this project documents and maps these diverse environments, in order to provide a fuller and more vivid depiction of the religious landscape of the early twentieth-century United States.With the generous support of the National Endowment for the Humanities, the project is currently digitizing approximately 232,000 schedules from the 1926 U.S. Census of Religious Bodies, a treasure trove of congregation- and place-specific data. These schedules will be made available on the project's website as photos of the records and as a transcribed dataset. The project is also investigating denominational records and other sources of data. Finally the project will use these datasets to create maps and visualizations which offer a rich depiction of how congregations related to one another in their local environments.

Death by Numbers (https://deathbynumbers.org). One of the most dreaded diseases in early modern England was plague, which was present in the British Isles from 1348 until 1679. The

most well-documented epidemics of the early modern era were in England's cities, particularly London, which suffered six major epidemics in the century between 1563 and 1665, and lost an estimated 225,000 people to plague. Government officials attempted to quantify the severity of various plague outbreaks and, starting in 1603, published London's weekly mortality statistics in broadside series known as the Bills of Mortality. The bills grew to include not just plague deaths but also dozens of other causes of death, such as childbirth, measles, syphilis, and suicide, ensuring their continued publication for decades after the final outbreak of plague in England. The weekly bills were also supplemented annually with a general account of the preceding year, published on the Thursday before Christmas. Between 1603 and 1752, almost 8,000 different weekly bills were published, chronicling plague and general mortality through the city of London. One of the major aims of the Death by Numbers project is to transcribe and publish the information in these bills in a dataset suitable for computational analysis.

Slave Voyages (https://www.slavevoyages.org). The SlaveVoyages website is a collaborative digital initiative that compiles and makes publicly accessible records of the largest slave trades in history. Search these records to learn about the broad origins and forced relocations of more than 12 million African people who were sent across the Atlantic in slave ships, and hundreds of thousands more who were trafficked within the Americas. Explore where they were taken, the numerous rebellions that occurred, the horrific loss of life during the voyages, the identities and nationalities of the perpetrators, and much more.

Appendix 2: Suggested readings

James Baker, "Preserving Your Research Data," Programming Historian 3 (2014), https://doi.org/10.46430/phen0039.

Julia Flanders and Trevor Muñoz, "An Introduction to Humanities Data Curation" https://archive.mith.umd.edu/dhcuration-guide/guide.dhcuration.org/intro/

Lincoln Mullen, "Introduction" Computational Historical Thinking https://dh-r.lincolnmullen.com/introduction.html

Christof Schöch, "Big? Smart? Clean? Messy? Data in the Humanities" Journal of Digital Humanities vol 2 no. 3 (Summer 2013), http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/