Electronic Currents

Assistant Editor: Sarah Dorpinghaus, University of Kentucky.
Contact Sarah at sarah.dorpinghaus@uky.edu if you would like to guest author a column or have a good idea to share.

Visualizing Archives and Library Collections

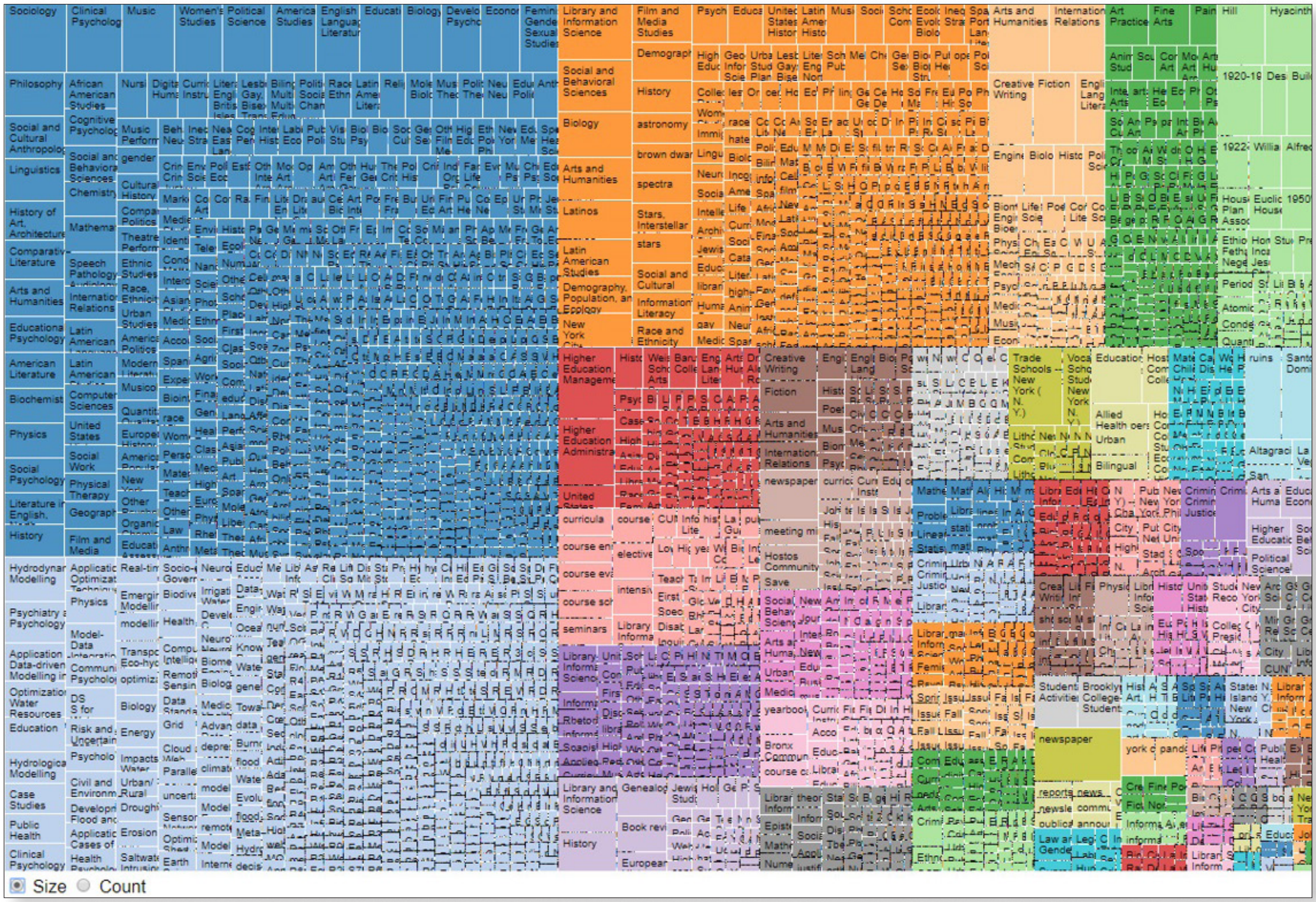
By Thomas Cleary, Archivist, LaGuardia Community College, CUNY

Introduction

Archivists and special collections librarians have struggled for a long time with how to show patrons what we have in our holdings. Collections have been made accessible through container lists, finding aids, and collection and content management systems such as ArchivesSpace, Islandora, and CONTENTdm. Each of these documents and systems also has its own learning curve and different functions, but even then the scale of some topics in collections or the connectedness between collections is not always apparent.

Here enters the world of data visualizations. Data visualization is a technique used for making data (in our case, EAD, MODS, and Dublin Core records) easier to understand in a visual format.

In my own time working in archives and libraries, I have been interested in using data visualization as both an access tool and as a way to analyze collections. As an access tool, visualizations can make it easier for visual learners to understand what is in the collections and how materials are connected. For analysis, visualizations let the archivist examine collections in a new way, possibly bringing up new topics or themes otherwise hidden in the metadata. These visualizations can also be used to show administrators what a collection specializes in and to make cases for developing different parts of collections. Overall, each of these purposes either increases access to the materials, or acts as a way to guide development and resources.



A visualization of the CUNY Academic Works repository made with the GLAMViz tool. Each color represents the contributions of a different college; each tile a different subject within the college's collection.

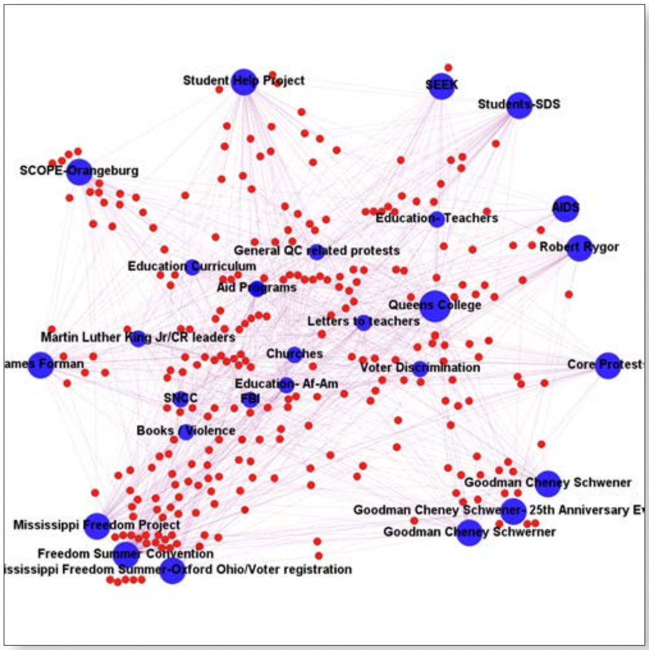
The GLAMViz Project

My most recent effort in making a visualization tool has been the GLAMViz project, which lets users analyze and view the different subject headings in their collections. GLAMViz is a digital humanities project coming out of a course at the CUNY Graduate Center's master of liberal arts program. Here I worked with two other students, John Parker and Carolyn Cea, to create a visualization tool aimed at people working in galleries, libraries, archives, and museums (the GLAM fields). We acknowledged that people in these fields have very different backgrounds, skills, and resources, but that those factors should not be a hurdle for making visualizations. With that in mind, we made the tool as simple to use as possible, automating each stage, but also programming it to create files at each step of the way. These files give users other opportunities to look into their data and modify them on their own. We decided to focus on visualizing subject headings as this could help patrons browse through collections, act as a tool to show students the focus of the archives, and give archivists a way to determine if their digital collections over- or underrepresent any specific parts of the archives.

Initially, we planned on creating an Islandora module that would display subject headings in a repository's data in a D3.js graph. I use an Islandora repository software at LaGuardia Community College to host our online collections, and I participate in the New York City Islandora working group, along with others interested in a tool like this. However, we quickly realized this project was too complicated to finish in a semester as none of us had any real coding experience.

Our change of direction led us to look at how we could use a Python script to harvest data from general OAI-PMH APIs and then have the script transform the data to work with a D3.js visualization. This proved to be much easier to complete in our timeframe and still let us achieve the goal of developing an easy tool to use. As the project proceeded, we added in a basic user interface that runs locally in the user's web browser. Finally, to allow end users to use the program without having to install Python and the environment, we created a virtual environment package using the Python venv module. This, paired with step-by-step instructions, allows just about anyone to use the tool.

While the tool has difficulty displaying larger repositories, it does work well for smaller repositories and successfully visualizes both the sizes of collections within a repository and which subjects show up the most. Our team plans to develop the project to offer different visualization templates and the choice to display fields other than subjects. More information about the project along with blog posts documenting its development can be found on our website at https://glamviz.commons.gc.cuny.edu. This website also has links to the Github page where the code rests. If you try out the tool, please feel free to get in touch and share your suggestions and comments.



A network diagram shows connections between items (red) and topics (blue) within the Queens College Civil Rights Archive.

Networking a Civil Rights Digital Collection

We based a different project on the digital collection of the Queens College Civil Rights Archive. The collection is hosted on an Omeka website and contained 325 objects at the time of the project in 2014. The goal was to use MALLET, a natural language processing classification tool, to "read" metadata and transcriptions of items copied from the website and then to take those topics and graph them as a network diagram. The network diagram would allow people to see how the materials clustered around specific topics. The final result was an online interactive network diagram that lets the archivist and patrons explore the connections between items, ideally improving reference and research around those materials.

(Continued on page 32)

(Continued from page 31)

Why all this effort to have a machine reclassify items that have already been fully described? The idea was to see if MALLET's natural language processing function would identify any topics unnoticed before because the materials were dispersed among collections and the archivist didn't notice the connection in the context of the collection. An algorithm reading the texts might also reveal that two items from different collections share connections with the same topic and might show the strength of the connection through a measured weight. This ability to add a numerical "weight" to the connections is a benefit MALLET has over regular description, and it enables a network diagram to be made.

The process was labor intensive, starting with making sure each object was properly described and conformed to the same standard. Also, each document with text was run through Adobe Acrobat Pro's OCR function to create transcriptions that could be saved as text files. The relatively small size of the digital collection made this fairly feasible to do manually. The metadata and transcriptions were combined into individual text files, each file representing an item, so they could be processed by MALLET. After a few runs, MALLET came up with 27 topics that I decided were relevant to the digital collection as a whole. The resulting data were put into Gephi, a network analysis software, and then graphed as a force diagram. The force diagram clustered the items around their most relevant topics, pulling the items that related most strongly closer to the topic. After moving clusters around for legibility, the final diagram was exported and synced up with the OII InteractiveViz tool, which lets Gephi diagrams be displayed on the web.

The resulting diagram can be found at https://archives.qc.cuny.edu/civilrights/topicweb. I found it useful to me as an archivist, as it shows how the collections group around a few single topics. A few overlooked topics and themes did become apparent, such as churches and education. While these two topics are common and fairly central, they serve more as linking themes, as items do not gravitate strongly to them. The stronger and more apparent topics appear on the outside of the diagram and show that items related to these topics vary strongly. I was unable to do any official surveys to see how researchers respond to the network diagram, but among the few I asked, the general consensus was "That looks cool, but how do I use it?" This feedback shows that further work needs to be done to make diagrams more understandable or to provide training on how to use them.

Further Readings

These two projects do not stand on their own, but related to a growing number of projects that use data visualization in archives and libraries. I found the following articles helpful.

1. Anne Bahde, "Conceptual Data Visualization in Archival Finding Aids: Preliminary User Responses," *portal: Libraries and the Academy* 17, no. 3 (2017), https://muse.jhu.edu/article/664309.
2. Mark Eaton, "Seeing Library Data: A Prototype Data Visualization Application for Librarians," *Journal of Web Librarianship* 11, no. 1 (2017), https://doi.org/10.1080/19322909.2016.1239236.
3. A. Miller, "Visualization Praxis: Data Visualizations with an Interdisciplinary Advantage," *Journal of Interactive Technology & Pedagogy* (February 7, 2017), https://jitp.commons.gc.cuny.edu/visualization-praxis-data-visualizations-with-an-interdisciplinary-advantage.