Collections As Data: Preservation to Access to Use to Impact

Short Paper

J. Mears
Library of Congress
United States
jame@loc.gov

A. Potter
Library of Congress
United States
abpo@loc.gov

K. Zwaard
Library of Congress
United States
kzwa@loc.gov

ABSTRACT

The Library of Congress has many millions of objects it has either digitized or acquired in born-digital formats. The first priority is to preserve, process, and provide access to digital collections via loc.gov. Most of these collections are free and available to anyone around the world, and their existence online represent the culmination of a huge amount of time, effort, and expertise. The Library of Congress has embarked upon an effort to maximize the use of our digital collections, promote innovation at the institution, and grow national capacity for cultural memory. This paper will describe how developing and executing a program around promoting collections as data has enabled the expansion of internal and external partners, deepened the exploration and value proposition of digital collections, engaged a broader community, and provided new skills to librarians who are working with collections as data. The programs that will be outlined will include the development of a digital scholarship lab which seeks to provide support for computational engagement with collections, testing different training models for digital skill building, and launching Innovator-in-Residence programs that showcase how data analysis techniques can change digital scholarship and digital curation landscapes. These partnerships and pilot projects build on the digitization and preservation efforts of the Library by broadening the scope of engagement and providing compelling use cases that amplify the impact of digital collections.

KEYWORDS

Data Analysis, Digital Scholarship, Outreach, Education & Training, Impact of Preservation

1 LAUNCHING NATIONAL DIGITAL INITIATIVES

The National Digital Initiatives division was founded in October 2016 and operates in the National and International Outreach service unit under National Programs. The core staff came from the retired National Digital Information Infrastructure and Preservation program after a realignment of the Library of Congress service units. The founding principle of NDI is to honor the knowledge, skills, and effort of the Library of Congress staff who build our digital library. Much of this is invisible work. The acquisition, transfer, processing, and preservation of petabytes of digital collections is central to the Library's mission and purpose and represent an incredible asset that belongs to the American public. NDI also inherits the network building and community engagement success of the NDIIPP program.

NDI's mission is to broaden awareness of the Library's innovation and use of its digital resources through outreach and external partnerships. We seek to work closely with other service units and divisions to reach new audiences, spark innovative projects, and work together on shared problems. To execute this strategy, three program areas were defined: 1) Expand the use of Library of Congress digital resources, 2) Incubate, encourage, and promote digital innovation, and 3) Grow national capacity for cultural memory.

2 ENABLING NEW USES OF THE DIGITAL COLLECTION

2.1 Collections as Data

The Library of Congress and other libraries have been serving digital collections online for over a decade. However, the model for which these resources are accessed largely recreates an analog experience of reading a page at a time or browsing through a collection of photographs. With modern computing power and the emergence of data-analysis tools, these digital collections can be explored more deeply and reveal more connections. In order to take advantage of computation, collections must be made available in a form that can be recognized by computers, a transformation not unlike moving from creating a card catalog to the MARC record schema. Obviously, the collection must be digitized but often more than a just a digitized image is needed to make full use of the object. Optical character recognition that reflects underlying text or metadata that describe the content in a structured way, like using ISO date and formats, latitude and longitude coordinates, or tags according to a taxonomy, are data that widely available tools can use to create visualizations and perform analysis.

For users that want to work with a large number of digital files and download them in bulk rather than one at a time, The Library of Congress provides processed bulk-data-download derivatives for two collections. The MARC Distribution Service provides all of The Library of Congress catalog records (https://www.loc.gov/cds/products/marcDist.php) and the National Digital Newspaper Program provides the raw OCR text of newspaper pages from the Chronicling America project (http://chroniclingamerica.loc.gov/ocr/). There are additional collections for which there is an approximation of bulk data access for anyone with the technical capacity for scripting and the aptitude to understand and crawl an API or a site. Affixing the text "?fo=json" to the end of many URLs in https://loc.gov results in a JSON API that can get some researchers started (see http://www.loc.gov/pictures/?fo=json). XML sitemaps are available for, currently, 277 digital collections (https://www.loc.gov/collections/sitemap), providing information such as the item URLs, last modified date, and frequency of updates (e.g. https://www.loc.gov/collections/franklin-pierce-papers/?c=1000&sp=1&fo=sitemap). Similarly, https://id.loc.gov has a number of formats available for crawling a large number of datasets, but the bulk download of data is not generally available.

The World Digital Library (https://api.wdl.org) has a well-documented API which serves a similar purpose for those collections. There are collections which do not have an API, documented or otherwise, that allows crawling or web harvesting. Some very motivated researchers have previously used this kind of interaction to build datasets, or to extract particular curated selections, and to make them available elsewhere.

Locations away from our main website occasionally have capabilities like one of those described above. For example, the Library of Congress participates in the Flickr Commons (https://www.flickr.com/commons) which comes with some capabilities of using the Flickr API for downloading collections.

These few examples represent a small percentage of the roughly 300 collections and approximately 2 million items, in multiple content formats, available on loc.gov but not optimized for anything other than interactive browsing and various degrees of searchability. One of the goals of NDI is to promote standards and practices around providing access to collections as data so that the full value of computation can be leveraged to bring even more awareness of the knowledge and creativity contained in the world's libraries.

2. 2 Exploring Digital Scholarship

Emerging disciplines — like data science, data journalism and digital humanities that take advantage of new computing tools and infrastructure — provide a model for creating new levels of access to library collections. Visualizing historical events and relationships on maps, with network diagrams and analysis of thousands of texts for the occurrence of words and phrases are a few examples of what's possible. NDI is actively exploring how to support these and other kinds of interactions with the Library's vast digital holdings.

Michelle Gallinger and Daniel Chudnov were asked by NDI to study how libraries and other research centers have developed services that use computational analysis, design and engagement to enable new kinds of discovery and outreach. Their report (PDF)[1], was just released. For the report, they interviewed researchers and managers of digital scholarship labs and worked with Library staff on a pilot project that demonstrated how the collections could be used in data analysis. This work resulted in concrete recommendations to the Library on how to approach setting up a Lab at the Library of Congress. These recommendations could also be helpful to other organizations who may be thinking of establishing their own centers for digital scholarship and engagement.

The recommendations from Gallinger and Chudnov would warrant a paper on its own but we will review some of their feature recommendations: First, that the Lab should be service oriented to outside users. Second, that the Lab support users to grow and develop their research. Third, that the Lab be an educational space, and fourth that the Lab enable organizational transformation. As research, scholarship, and resources become more digitally based, the services provided by the Lab will, in the beginning, be groundbreaking but eventually become part of the normal services offered by the Library. It is this transformation of how librarians perform reference services for the digital collections—services that include identifying available data, working with that data to deliver specific a specific corpus, establishing provenance for the data, and helping to answer research questions with the data—that NDI hopes to cultivate in a digital scholars lab.

Concurrently, NDI is co-leading an internal group charged with studying how the Library of Congress can programmatically enable digital scholarship for its collections. The group is completing a paper in 2017 that describes the challenges and opportunities inherent in the Library of Congress' collections and operations. Much of the information in the Collections as Data section of this paper comes from that group's efforts.

2.3 Engaging Communities

Using Gallinger and Chudnov's report and recommendations as guidance, an on-site digital exploration center is an opportunity to open the Library's digital holdings up for enhanced scholarly inquiry. It would also serve as a point of engagement for more general users for topic or visual exploration It would serve numerous communities:

- Scholars interested in using Library of Congress digital collections as data to create visualizations, explore digital scholarship, and pursue computationally-assisted research.

- Teachers and students who want to use Library of Congress data in their STEM or humanities classwork and projects.
- Artists, writers and other creatives who want to access our public domain content to remix and reuse the Library's digital materials in new ways.
- On-site and virtual users who want bulk access to digital collection data for analysis or use in other applications.

In developing coordinating services, the Library will work collaboratively to design orientation and introductory classes on digital scholarship methods and tools to help introduce the concept of using library collections as data. The courses will be in parallel to professional development trainings for Library staff who will be delivering services, especially reference services, in new ways.

2.4 Skill-building

Providing access to digital library collections as data requires new skills for librarians and using digital library collections similarly requires new skills and modes for researchers. NDI is exploring how to offer educational opportunities for both of these groups to gain these new skills. In February 2017, 40 librarians, archivists and data wranglers came to the Library of Congress to learn advanced skills in managing digital collections. The Library hosted the Software Carpentry workshop, inviting staff from the institution, the DC Public Library and other federal libraries for hands-on learning in the programming language Python, the version-control software Git and the command-line interface Bash. Software Carpentry is a volunteer, non-profit organization that provides short, intensive workshops to help researchers automate tasks and manage information. It started with scholars in the physical sciences who found that traditional graduate programs were not preparing them for the challenges of working with data for their research products. Software Carpentry workshops have lately been adapted for social sciences, the humanities and libraries.

Attendees from the workshop immediately saw a way to apply their new skills. “I can see an opportunity to use scripts to improve researchers' experience in the reading room,” said Kathleen O'Neill, a senior archives specialist in the Manuscript Division. “For those researchers with limited experience with digital collection material, we could provide a library of simple scripts to search, analyze and report on the born-digital collection material.”

Later in the year, NDI partnered with George Mason and George Washington University Libraries to host a program titled “Hack-to-Learn”. There is clear demand for hands-on computational experience yet librarians are often under represented at events like hackathons. So the organizers worked together to develop an inclusive hackathon that utilizes the skills librarians already have and introduces them to low or no-cost computational tools to explore digital library collection data sets. Over two days, 61 librarians and researchers attended, most with no prior programming experience. They were given instruction around MALLET, a topic modeling tool, Gephi, a network analysis visualization tool, OpenRefine, a data editing tool, and Carto, a mapping tool. According to post-event surveys, attendees confirmed that the event was valuable in orienting them to computational methodologies and new research and processing possibilities for their digital collections.

Developing and promoting a program to develop skills around collections as data has enabled the expansion of internal and external partners, deepened the exploration and value proposition of digital collections, engaged a broader community, and provided new skills to librarians who are working with collections as data.

3 SHOWCASING DATA ANALYSIS TECHNIQUES

3.1 Innovators in Residence

While 3 million people access the collection every month on the loc.gov website, there is untapped potential of using the digital collections in other ways, be it for data analysis to support digital scholarship, by an artist seeking to remix and reuse content to create new art or commerce, or a journalist in need of authoritative data for in-depth reporting.

NDI has launched an Innovator-in-Residence program to support innovative uses of our collections and partnerships with universities, scholarly societies, artists, corporations, and other organizations. A wide variety of programs would be included in a broad residency program:

- A program that would support the work of digital humanities scholars at the Library of Congress, which have been successful at other large libraries.
- A challenge grant program that would run contests for innovative uses of digital collections and make small awards to winners, modeled after Federal Data Challenges.
- A program that would support software developers to come to the Library and build tools and services that make innovative uses of Library collections.
- A fellowship program to support hybrid teams to use Library of Congress collections and data to support information for the American people. A journalist and a social scientist working with a historian working on a project that will be published widely in a newspaper, magazine or book.
- A program to support artist to create art based on the Library's collection.

In 2017 two multimedia artists were selected to create digital art pieces based on the Library of Congress digital collections. In addition to creating the works, they will present their art to the public and Library staff in a workshop.

3.2 Crowdsourcing application

NDI piloted the Innovator-in-Residence program at the Library with an opportunity for staff to apply for a short-term assignment. Two staff members were selected. One, Chris Adams, focused on exploring automatic image identification and fleshing out the concept of labs.loc.gov. The other, Tong Wang, created a proof-of-concept crowdsourcing application. Tong wanted to show how we could leverage open source tools created by other members of our community (specifically, the Scribe http://scribeproject.github.io/ transcription framework) and the Chronicling America API (http://chroniclingamerica.loc.gov/about/api/) to engage the public with library collections and to enhance the usability of our digital material. The application he created is called 1000 Words and it invites users to identify cartoons, photographs, and other images in historic newspapers and to describe them. This would allow those images to be searched for the first time and for data sets to be created. For example, the Library could offer to researchers a collection World-War-I-era cartoons, which previously would require flipping through thousands of newspaper pages to collect.

3.3 Events

We have hosted several events in an effort to demonstrate what we mean when we talk about collections as data and provoke exploration of our collections computationally for our staff and members of the public.

We invited programmers, artists, entrepreneurs, activists, researchers and librarians finding new ways to connect with digital collections to share best practices and lessons learned during *Collections as Data: Stewardship and Use Models to Enhance Access*, a public conference held at the Library of Congress in September 2016. From tracking the popularity of bible verses in historic newspapers to mapping everywhere you've ever walked on your smartphone, speakers such as Jer Thorp from the Office for Creative Research, Pinboard founder Maciej Ceglowski, and Director of the Five College Digital Humanities Initiative Marisa Parham discussed the methods, skills, ethical considerations and community participation necessary to find new meaning in digitized and born digital cultural heritage collections. The topic resonated with many. The event's hashtag #AsData trended #2 on Twitter the day of the event, receiving over 8500 livestream views. The more than 450 people came, representing over 15 different countries from universities, public libraries, galleries, federal institutions, and newspapers such as the Washington Post and New York Times. Tweets and verbal feedback revealed attendees appreciated the holistic, nuanced, cross-disciplinary consideration of data.

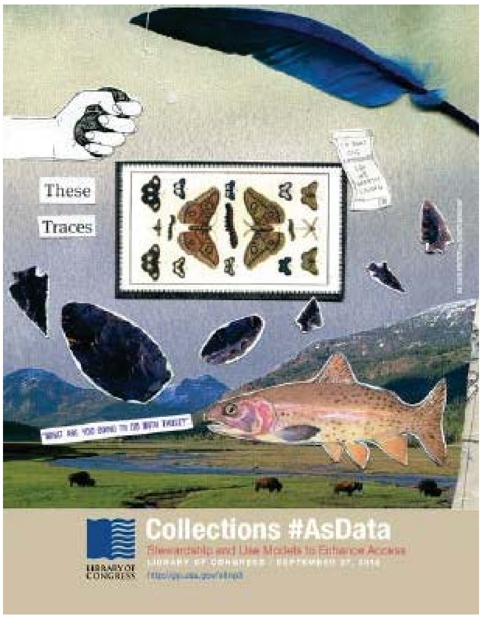
These
Traces
"WHAT ARE YOU GOING TO DO WITH THAT?"
Collections #AsData
Stewardship and Use Models to Enhance Access
LIBRARY OF CONGRESS / SEPTEMBER 27, 2016
LIBRARY OF CONGRESS

Figure 2: #AsData poster designed by Oliver Baez Bendorf.

The program was so successful that we will host a second *Collections as Data* Symposium in July, showcasing the impact of computational research through a series of successful case studies across disciplines and practitioners.

To facilitate learning and discovery, we were the host site for Archives Unleashed 2.0 in 2016, a web archive datathon series facilitated by Matthew Weber, Ian Milligan and Jimmy Lin in which teams of researchers used a variety of analytical tools to query web-archive data sets in the hopes of discovering some intriguing insights before their 48-hour deadline is up. Aside from hosting, the Library of Congress participated on the Hackathon teams and by providing collections for the teams to work with.

In an effort to centralize participatory opportunities for local District of Columbia residents, develop best practices for community engagement, and support sharing of information across DC cultural heritage institutions, NDI is co-hosting a Citizen Public History Fair in September with NARA, the Smithsonian Transcription Center, DC Public Library Special Collections, United States Holocaust Memorial Museum, the Folger Shakespeare Library, the Lincoln Theatre and the Anacostia Community Museum. The fair will showcase or demo volunteer project opportunities including digital crowdsourcing initiatives and feature lightning talks from curators and researchers. Following the event, a series of crowdsourcing programs by our institutions will take place throughout the year in an effort to develop a robust community of practice around historical resources.

In 2018, NDI will also serve as co-host of several high-profile meetings at the Library of Congress that support the collections as data vision. These include the 2018 Code4Lib Annual meeting, the 2018 IIIF Annual meeting, a HathiTrust Research Center workshop, and a workshop in conjunction with a grant funded DARIAH-EU project on digital heritage and culture.

4 CONCLUSION

The Library of Congress has millions of digital items available to the American Public via the loc.gov web site. Millions of people from all over the world visit our vast collections online. An additional 1.6 million people visit the Library of Congress Jefferson building each year [2]. The efforts of the National Digital Initiatives to support the practice of providing data level access to digital collections provides opportunities to enhance the preservability of digital collections and increase their usefulness to the communities we serve. When digital collections are leveraged as data, the public spaces, technical environment, training, tools and expert support are all modernized and made more relevant. These partnerships and pilot projects described above build on the digitization and preservation efforts of the Library by broadening the scope of engagement and providing compelling use cases that amplify the impact of digital collections.

REFERENCES

[1] D. Chudnov and M. Gallinger. 2016. *Library of Congress Lab: Library of Congress Digital Scholars Lab Pilot Project Report*. Washington, DC.

[2] OCIO Web Metrics & Analytics Service Team. Library of Congress.2016. *Quarterly Key Metrics*. http://staff.loc.gov/sites/webmetrics/reports/