Library of Congress Lab

Library of Congress Digital Scholars Lab Pilot Project Report

The Lab envisioned by the authors in this report is based on successful models that support access and use of digital collections, which have been drawn from interviews with experts and scholars and shaped by practical experience. The recommendations made in this report consider the Lab as part of the unique structure and offerings of the Library of Congress.

Michelle Gallinger and Daniel Chudnov

12/21/16

LIBRARY OF CONGRESS LAB

Library of Congress Digital Scholars Lab Pilot Project Report

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The National Digital Initiative (NDI) team at the Library of Congress hired Daniel Chudnov of Chudnov Consulting and Michelle Gallinger of Gallinger Consulting to explore how to deliver Library of Congress digital collections as data to researchers. The demonstrated need of the Kluge scholars for more assistance in their work with digital collections inspired the work. The authors drew upon their own experience as well as interviewed researchers, subject matter experts, and worked with Library staff to perform the transfer pilot. The insights from these activities shape the report. The authors make recommendations for how the Library of Congress might best establish and grow a Lab. These recommendations are the opinions of the authors on what the Library could and should do to ensure a successful endeavor.

The Library of Congress Lab envisioned in this report is modeled after successes in emerging digital scholars labs, digital humanities centers, digital reference services, and other departments dedicated to supporting access and use of digital collections. The recommendations made in this report shape the Lab in response to the unique structure and offerings of the Library of Congress. The Lab provides reading rooms, reference staff, and curatorial staff with technical support to allow users unprecedented access to digital collections. The Lab supports resident scholars and fellows in their attempts to engage with digital scholarship. The Lab connects users to the unparalleled reference staff of the Library for content expertise. The Lab helps Library staff to consider cross-departmental workflows, organization, structure, and internal staff development at a time of great change driven by rapid development of digital scholarship and infrastructure. The Lab hosts experiments, seeks and highlights creative external partnerships, and is a laboratory for testing new ideas.

The Lab serves the Library of Congress in three main capacities:

1. ***It responds to the demonstrated need expressed by Kluge scholars and other researchers and users who desire increased access to and guidance on how to use the Library's unique digital collections.***

2. ***It would support existing Library of Congress reference and curatorial staff. The reference and curatorial staff would continue as primary experts in their areas with the Lab functioning as a support for them in fulfilling user requests for digital collections.***

3. ***It would function as a laboratory in which the Library could test out transformative services. Successes from the Lab could then be reintegrated into the Library as they mature.***

It is now possible for the Library of Congress to provide transformational access to its digital collections. A Lab at the Library of Congress is an opportunity to build on institutional strengths and to answer the challenges of serving evolving models of 21st century

scholarship. In developing and offering a Lab, the Library will extend its existing collection specialties, leverage its existing digital collections, and through collaboration, maximize impact of its work while avoiding overlap with other institutions. The Lab will enable the Library of Congress to build a community around providing access to and developing insights from digital collections.

Background

The Library of Congress has historically provided reference support to users as they use the collections through the reading rooms. As the Library of Congress acquires more digital collections, there is a need to extend this tradition to support researchers, scholars, and other users interested in the content in these digital collections as data, often at larger scales than before, and often with new methods and computing techniques. As user requirements for access and manipulation of digital collections changes, so too must the services offered by the Library of Congress in support of the researcher.

About This Report

In August 2016, The National Digital Initiatives (NDI) division at the Library of Congress, in partnership with the Library's John W. Kluge Center, hired Daniel Chudnov of Chudnov Consulting and Michelle Gallinger of Gallinger Consulting to explore how to deliver Library of Congress digital collections as data to researchers. The desire to provide enhanced support for digital scholarship for the residential scholars of the Kluge Center inspired this solicitation.

The contract included two major deliverables that are both addressed in this report:

1. A pilot demonstration showing a feasible path for delivering a set of Library collections as data to an on-site researcher, and
2. A report recommending how the Library could develop the methods, policies, and technical environment to facilitate advanced computational research with its digital holdings, with specific focus paid to residential scholars at the Kluge Center.

Defining the Lab

There is a wide spectrum of services and situations that different institutions call labs, digital scholarship centers, digital humanities centers, humanities computing environments, digital research services, and a variety of other names. In some cases they may be a physical space, or they may not; they may involve temporary funding and staff assignments, or permanent; they may enable scholarly investigation, process experimentation, pattern recognition, or any of a variety of activities we tend to think of as occurring within traditional life and social science laboratories within research centers, universities, corporations, government, and non-governmental organizations. Within this spectrum there are broad choices that will impact the vision, services and objectives of a Library of Congress Lab. Should a Lab house a physical space for scholars, staff, or both? Should the primary function of a Lab support scholarship, or organizational experimentation? Should a Lab be a permanent fixture or a short-term fulcrum for innovation? These questions of what a Lab should look like and how a Lab might be structured within the organization must follow from a clear and shared understanding of why a Lab is necessary.

The Library of Congress has been responding to the transformation in content production from traditional physical materials to digital content in a number of ways for many years. The Library has made bold acquisitions decisions such as taking the Twitter archive and actively selecting and crawling the web to create the Web Archives. The Library of Congress has been dedicated to preserving the digital content it acquires. It is now possible for the Library of Congress to provide transformational access to these collections as well. The Lab can provide the reading rooms and the reference and curatorial staff with technical support to allow users unprecedented access to the digital collections. In addition, the Lab can be designed to help the Library to consider cross-departmental workflows, organization, structure, and internal staff development at a time of great change driven by rapid development of digital scholarship and infrastructure. The Lab can host experiments, seek and highlight creative external partnerships, and allow for a safe space to serve as a testbed for new ideas.

IT IS NOW POSSIBLE FOR THE LIBRARY OF CONGRESS TO PROVIDE TRANSFORMATIONAL ACCESS TO ITS DIGITAL COLLECTIONS

Conceiving of a Lab as a service environment is in keeping with how digital scholarship centers are being defined as they are formed. In fact, "a key attribute that distinguishes digital scholarship centers from more traditional research institutes is that they are service organizations, staffed by individuals with specialized skills who support work in the digital environment."[1] The Library of Congress Lab needs to be defined as a service center providing offerings that support the work of researchers, scholars, and other users across all the digital collections of the Library. It must also be a service unit for the Library itself, providing training, reference support, tool review and support, as well as a forum for visiting scholars and external partners to exchange news and innovative developments in digital scholarship.

A Lab at the Library of Congress is an opportunity to build on institutional strengths and directly answer the challenges of serving evolving models of 21st century scholarship. As they are formed anywhere,

> "[d]igital scholarship centers can build institutional capacity to address emerging and future scholarship needs. The rationale for developing such a center includes to:
>
> - provide a support mechanism for the growing areas of e-research and digital scholarship;
> - bring together expensive technologies (and services to support their use) to serve the entire campus; and
> - let students explore digital technologies in their research when such resources — whether technologies or advisory support — are unavailable in their departments."[2]

[1] Educause case study. Joan Lippincott, Harriette Hemmasi, Vivian Lewis. *Trends in Digital Scholarship Centers.* Web. 30 November 2016. http://er.educause.edu/articles/2014/6/trends-in-digital-scholarship-centers

[2] Educause case study. Joan Lippincott, Harriette Hemmasi, Vivian Lewis. *Trends in Digital Scholarship Centers.* Web. 30 November 2016. http://er.educause.edu/articles/2014/6/trends-in-digital-scholarship-centers

Digital Scholars Labs have been and are being formed to serve research and scholarship of digital content. Most Digital Scholars Centers offer "a wide variety of services, including workshops and classes, opportunities and spaces for collaboration, support for digital pedagogy and instructional design, and access to various technologies."[3] For example, the Digital Scholars Lab at Brown University was "designed specifically for collaboration, flexibility, and ease of use for scholars working on data-rich and visually-mediated research."[4] At the University of Virginia the Scholars' Lab has been a space dedicated to advancing "digital scholarship in humanities, information and library science, social sciences, and related fields, emphasizing interpretative and theoretical as well as technological innovation and inquiry."[5] While there are elements that university libraries and digital scholars labs feature that are relevant, the Library of Congress is different from a university library and a lab at the Library of Congress will be most useful if it serves all elements of the institution as well as the needs of researchers and scholars.

A Lab at the Library of Congress

The Library Lab should explore how scholars use collections as data to further their research and scholarship and the various ways that research can be supported by the Library of Congress. The Lab should help humanize the Library; although it is natural and exciting that Library of Congress representatives introduce collections to guests with facts and figures highlighting the enormity of holdings, this sheer scale can intimidate first-time visitors. At the September 2016 Collections as Data event, one participant reflected on this concern, indicating that even as an experienced researcher, being presented with the overwhelming volume of material can feel like an obstacle to basic orientation and discovery within Library collections. The Lab can work to increase approachability of collections, whether a scholar might requires access to one item, one thousand items, or one million items. The staff of the Lab should play a key role in ensuring that every individual working with digital content has an interface to collections they can comprehend, and in offering support that fits research needs.

The Library has an opportunity to develop a Lab and enter a still-emerging space as a leader in digital scholarship collaborations. A Lab is an opportunity for the Library to create a program in which cross-disciplinary, multi-institutional, collaborative work can happen, be housed, and have a role as part of the Library's expanding digital collections. In developing and offering a Lab, the Library will extend its existing collection specialties, leverage its existing digital collections, and through collaboration, maximize impact while avoiding overlap with other institutions.

---

[3] Coalition for Networked Information (CNI). *Report Of A CNI-ARL Workshop Planning A Digital Scholarship Center 2016*. 2016. Web. 3 Sept. 2016. https://www.cni.org/wp-content/uploads/2016/08/report-DSCW16.pdf

[4] Web. 30 November 2016. http://library.brown.edu/dsl/

[5] Web. 30 November 2016. http://scholarslab.org/uncategorized/whats-the-scholars-lab/

A Lab can be a service for Library of Congress users, staff, and to the Library as a whole. One key service of the Lab would be a complete, exhaustive website detailing all bulk data collections, APIs, and services (such as IIIF, feeds, etc.), and experimental interfaces available from LC. This service would meet several goals at once: it can serve as a starting point for researchers seeking to dig into data; it can be a jumping off point for teachers and students looking to get started with methods and tools for digital inquiry; it can also be a focal point for LC staff, facilitating a shared understanding of what might be possible with digital scholarship.

The Lab should also be a center of service to LC staff as well as to external researchers. The internal relationships developed through this service are critical to the success of the Lab. LC reference staff are likely to remain the first line of support for external researchers looking to work with LC digital holdings. For the Lab to serve these researchers, the Lab must also serve the reference staff and have a strong relationship so that reference staff are aware of new offerings and significant opportunities from the Lab.

A Library of Congress Lab should be designed as a service to the Library as a whole as well as to the researchers, scholars, artists, and others looking to use the impressive digital collections of the Library. It should be conceived of as a transformational tool for the overall organization of the Library. While the Library of Congress has an impressive record in acquiring and preserving digital content, that digital content has not always been fully available to users. A Lab is an opportunity for the Library to address these challenges. It is also an opportunity for the Library to identify new techniques, workflows, and practices that can benefit all divisions, reading rooms, and teams. The Lab should be conceived as an incubator in which to determine what significant integration of digital content might look like in the entire Library. Once the Lab has identified particular kinds of change as successful or worthy of pursuit, it should involve staff from throughout the Library to implement and refine change. A Lab is also an opportunity for the Library to focus on continuing education in computational research methods that could be used to enhance existing networks with State Centers for the Book, Federal Libraries, National Digital Stewardship Residents, and Kluge scholars.

THE LAB SHOULD BE CONCEIVED OF AS A TRANSFORMATIONAL TOOL FOR THE OVERALL ORGANIZATION OF THE LIBRARY

A Library of Congress Lab would serve all scholars, researchers, and others interested in using Library of Congress digital collections in the humanities, social sciences, and life sciences, including audio, visual, geospatial, and other materials. In doing so, the Lab will increase the profile of the collections and increase digital scholarship at the Library of Congress. All the digital holdings of the Library would be supported by a Lab. Increasing the accessibility and profile of the Library's digital collections would benefit Kluge scholars. It would increase the use of the digital collections in the American Folklife Collection, Prints and Photographs, Geography and Maps, the Web Archive collections, the Chronicling America collection, digital holdings within the Motion Picture Broadcasting and Recorded Sound division, as well

THE LAB WILL GIVE THE LIBRARY INSIGHT INTO WHAT THE FUTURE DEMANDS ON THE LIBRARY AS A WHOLE WILL BE AS IT SEEKS TO SERVE THOSE NEEDS TODAY.

as many others. The Lab will give the Library insight into what the future demands on the Library as a whole will be as it seeks to serve those needs today.

Clear Goals in Line with Institutional Objectives

A Lab at the Library of Congress must be integrated into the Library of Congress core mission and programs and must offer assistance to the researcher using digital content held in any of the Library's many divisions and reading rooms. In the formation of a Digital Scholars Lab, "a clear definition of the center's goals is critical, especially as they relate to the broader institutional objectives. Sharp focus on direction and purpose will help garner support, as well as serve to guide operations."[6] The goals of the Lab should include to:

- ***Increase awareness and appreciation of Library of Congress digital collections so they may be used to their fullest potential.***
- ***Increase scholarship using Library of Congress digital collections, including providing patrons with access to collections as citable data sets so they can develop new insights.***
- ***Provide technological support to researchers and scholars using Library of Congress collections.***
- ***Provide technological support in concert with the subject matter expertise offered by the reading rooms and curatorial staff to help researchers employ digital content in ways that support their research and scholarship goals.***
- ***Increase technical staff expertise within the Library of Congress to serve and support patrons with data-related needs.***
- ***Increase the range of partnerships with peer institutions.***
- ***Increase the opportunities for on-site scholars to engage with digital collections in complex and meaningful ways.***
- ***Provide instruction and access to data scholarship methods and techniques to students, researchers, staff, and others interested in working with the Library of Congress digital collections.***
- ***Provide feedback to the entire Library with mechanisms for staff to identify and incorporate significant successes from the Lab environment into the Library as a whole.***

These goals are intended to support the Library of Congress institutional goals and strategic plan. It is imperative that the Lab support the organizational efforts and align with the Library strategic goals. Some ways in which the Lab can strengthen the overall institutional efforts include:

[6] Coalition for Networked Information (CNI). *Report Of A CNI-ARL Workshop Planning A Digital Scholarship Center 2016*. 2016. Web. 3 Sept. 2016. https://www.cni.org/wp-content/uploads/2016/08/report-DSCW16.pdf

- ***The Library of Congress continues to support and promote intellectual and creative endeavors. The Lab will support this work by offering a new forum in which ground-breaking work can take place.***
- ***The Library of Congress is a national leader. The Lab helps the Library lead in digital scholarship and collaborations.***
- ***The Library of Congress strives to acquire, preserve, and provide access to a universal collection of knowledge. The Lab identifies and refines methods for interacting with digital content as user needs continue to evolve. It also provides the Library of Congress an opportunity to incorporate new methods of access -- and the insights derived from that access -- back into the record it keeps of America's creativity.***
- ***The Library has a significant ability to act as convener and promote collaborations in innovative and emerging areas. The Lab supports the Library's efforts in collaboration.***
- ***The Library constantly seeks to organize itself in the most effective manner to fulfill its mission and goals for Congress and the American people. The Lab will support the Library in facilitating change within the organization. It will act as an incubator where new approaches for access and collaboration can be safely tested. Once determined to be effective, the Lab can work to integrate proven methods into the practices of the entire Library organization.***

For a Library of Congress Lab to be successful, it must support digital collections, the curatorial staff, the reference staff, and the researchers of the entire Library. All collections available to the public on-site at the Library must be supported by the Lab, with Lab staff supporting existing Library Services efforts to help researchers. Curatorial and reference staff could funnel queries from Ask A Librarian that need additional technical support to the Lab.[7] Lab staff could then work with Library Services staff to serve the researchers. In this way, the Lab must be well integrated into the fabric of the Library. It is critical that the staff of the Lab build strong relationships with the Library of Congress reference and curatorial

---

[7] The Library is already getting these questions. Chronicling America Data Questions from April 2016-December 2016 via the ChronAm-Users Listserv **(CHRONAM-USERS@LISTSERV.LOC.GOV)** include:

1. "If there's an option somewhere that could let me just get the page URLs for my query results (see example below), that would let my app retrieve large result sets from Chronicling America about 50 times faster than the standard result page JSON. Anybody know of some way to request that?"

2. "Would anybody happen to know whether there exists a representation of how Chronicling America's collections are distributed geographically and/or over time?
I'm working on researching how a particular phrase is distributed in the newspapers (yes, I realize there are all kinds of problems, but just thought I'd use this as rough approximation for whatever it might be worth). So I'm plotting on a map how many hits of that phrase I found in Chronicling America newspapers. But of course that is influenced at least as much by what the database covers, so I'd like to at least weight it by what's in the database. That is, did the phrase occur in some particular location more often than would be warranted by how much material is digitized from that location.
I could, I suppose, pick a common phrase and download all the pages that have that phrase, and then extract the location information from that. But I'm wondering if there's a better, less time-consuming way?"

staff. Reference and curatorial staff will come to refer relevant queries to the Lab when they have a confident understanding of its offerings, availability, and staff expertise. Ongoing communication and relationships are necessary to foster that exchange. The Lab will require some measure of autonomy in providing tools and services to scholars. Local computing capacity, or cloud-hosted resources, for example, could be prepared for short-term use by specific scholars in concert with particular collections in a secure manner.

While supporting the entire Library's digital collections and curatorial and reference staff is key to long term success of the Lab, identifying strategic partnerships and focusing on developing strong relationships that provide a model of what the Lab can do are important immediate steps. The Lab should focus on two or three key partnerships within the Library where it can develop and maintain significant relationships. The Lab could explore how it can support the John W. Kluge Center by offering digital scholarship introductions to new fellows, provide a list of recommended digital scholarship tools and examples of they might be best employed, and help Kluge fellows connect with the digital content available at the Library of Congress. Another near-term partnership could be for the Lab to engage with Teaching with Primary Sources (TPS) to support the use of Library of Congress collections as data. The Lab could provide thoughtfully packaged bulk downloads of Library of Congress collections for K-12 students to engage with, and in so doing, promote emerging digital scholarship capabilities and innovative thinking in younger users.

Developing the Lab as part of the National and International Outreach office (NIO) at the Library will help the Lab perform key outreach and relationship building necessary to the success of the endeavor. The Lab needs to engage with a community of practitioners who are still developing, refining, and inventing digital scholarship practices. While some techniques are well established, such as using digital collections to study transportation patterns with network analysis, and using text mining to study how language changes over time, others techniques continue to emerge and be applied to new collection types and domains. NIO can help the Lab to connect to other similar efforts nationally and internationally, promote exchanges of practice and effort, and encourage the widest possible use of Library collections and Lab services.

The Library Lab extends the ability of the Library of Congress to build a community around providing access to and developing insights from digital collections. The Library has significant relationships with peer institutions near and far, such as the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland, College Park, the British Library and its Labs, and many other organizations doing significant work in providing access to digital collections and support to those seeking to use them. These relationships should be cultivated and maintained. At the same time, while it is critical for the Lab to connect with significant players in digital scholarship, digital humanities, and digital computing, the Lab also offers the Library an opportunity to increase the diversity of scholars, researchers, and communities engaging with the Library's digital collections. It is imperative to be mindful of underrepresented groups, research projects, and collections when reaching out and building a community of collaborators and users for the Lab. This opportunity should be capitalized on and significantly emphasized in the creation and ongoing efforts of the Lab.

LIBRARY OF CONGRESS LAB PILOT ACTIVITIES

As part of developing this report, the authors were tasked with managing a pilot activity to demonstrate the feasibility of delivering a set of Library collections as data to an on-site researcher. To round out this work, we interviewed a number of researchers and subject matter experts, recommended a data set from among Library digital collections not currently available to the public, and worked with Library staff to transfer selected content to a secure third-party hosting service, where we performed format transformation, information extraction, and summary analysis steps to proxy both the potential roles of the Lab and researchers using Lab services. In this section we summarize the technical pilot activity and identify several lessons learned from the process.

Technical Pilot

In the pilot, we identified a digital collection that met particular criteria, worked with Library staff to transfer part of the collection to a secure third-party computing platform, and simulated the activities of format conversion and basic analysis via processing at scale. These are activities likely to be part of many workflows developed by a Lab on behalf of scholars. From this successful pilot we gained a number of insights, which directly informed our recommendations.

Content Selection and Transfer

To focus our efforts, we sought collections that were digital, had clear access restrictions, and that allowed use in some capacity. Additionally, we were interested in collections that presented significant challenges to moving and working with the data. In particular, we considered collections with features such as:

- Large collection scale that would pose challenges in selecting and moving portions of it for scholarly processing (i.e. larger than could fit on a thumb drive or typical laptop)
- Collections with multiple media types (text, image, sound, moving image, etc.) that might pose processing challenges
- Collections containing multiple file format types
- Collections with available metadata
- Collections under the stewardship of curators with long-term responsibility to the collections who could help provide insight

Many digital collections at the Library of Congress fit these criteria. Ultimately, through guidance from Library of Congress staff, we chose to work with a subset of the Web Archives collection. This selection offered all of the features listed above and also allowed us to consider a file format (WARC) requiring specialized processing and which has met with growing scholarly interest in analysis using contemporary tools for distributed processing of

large-scale datasets. Additionally, this scholarly interest in studying terabyte-scale collections of archived web sites is not served fully by the prevailing standard for access to web archives, the Wayback Machine. The Wayback software, developed first at the Internet Archive and now used at the Library of Congress and elsewhere to host web archives on the web, allows access to the history of websites by clicking through pages as they appeared when captured in the past by crawlers. This access mode is useful, fun, widely admired, and meets some user needs for access to digital collections, but not all user needs for access to web archives. Browsing through the Library of Congress web archive collections via Wayback lets us see individual pages of handfuls of web sites at distinct points in time, but this one-link-at-a-time paradigm is a mismatch for the capacity of a historian armed with a computing cluster wishing to perform content or network analysis over millions of web pages containing billions of links. At the Library of Congress, many web archives collections are available online through Wayback, yet many more are stored on tape or are otherwise not yet queued up for access due to a variety of reasons. These reasons include a backlog of materials ready to go online, the cost of disk storage, and the competing priorities on a limited staff. Additionally, the Wayback interface itself does not provide bulk access to complete WARC files as data or to other web archive file formats aside from APIs to several data availability queries. Given these issues and the increasing value of access to web archives collections as data and at scale not otherwise being served currently by the Library of Congress, we believe this selection of web archive materials illustrates the gap in access to digital collections which a Lab might be positioned to fill.

To that end, we successfully transferred over five terabytes of WARC files with web pages collected from links in syndication feeds from major news sites the Huffington Post and the Detroit Free Press, comprising material including HTML, images, web scripts and style documents, and linked videos from YouTube which the Library collected in support of its broader collections. Because this particular set of materials were crawled with the specific intent to complement broader crawls, rather than to stand on their own as collections for public access, they are not otherwise scheduled for cataloging or public access and have not been studied on their own in any way. They have been collected on a daily schedule for over a year, allowing us to proxy a scholarly role by looking at patterns in how the materials evolve over this time period.

We selected Amazon Web Services (AWS) as our secure third-party hosting platform, and it worked well for the pilot. In particular, we used the AWS Simple Storage Service (S3) storage solution, Elastic Compute Cloud (EC2) for virtual computing, Identity and Access Management (IAM) for managing users with restricted access to materials and securing unique keys for each user in the process, and the Elastic MapReduce (EMR) distributed computing toolkit for computing jobs requiring substantial computing power from multiple machines. The choice of AWS as a platform was arbitrary, based primarily on familiarity and experience. Similar commercial offerings from Google and Microsoft, among others, are competitive alternatives with which we might have found similar success.

Our first challenge after selecting the collection materials was to transfer those materials securely to our prototype user. Through an AWS account we established, we created a storage location in which to store materials to be transferred from the Library of Congress Web Archives collection to the cloud. We configured this location to restrict access to read,

write, and even list contents to only specific users via an additional AWS account created strictly for Library of Congress staff to use, and another strictly for the consultant to use. These two accounts have differential rights; for example, the Library of Congress staff account has write access, but cannot delete materials, whereas the consultant account is configured with the ability to do both, as well as to read the contents of files for further processing. We tested these permissions on both accounts using the Library of Congress-developed "bagger-js" web-based application for transferring content in Bagit bags from local storage to AWS S3, a positive side benefit of which was hands-on testing of bagger-js. A small number of user interface issues discovered during the pilot process resulted in tickets written against the bagger-js GitHub repository, and hopefully will inform future development of enhancements.

To provide access to the AWS Library of Congress staff account to Library of Congress staff, we delivered a pair of keys generated by IAM through a separate third-party service trust management service called Keybase. We encrypted the key pair using Keybase and emailed them, in encrypted form, to Library staff, who in turn received the encrypted keys, decrypted them using Keybase, and was then immediately able to use the keys to transfer materials to the specified location using the free AWS command line tool. After a successful initial test, we proceeded to transfer 17 bags successfully, each containing hundreds of up-to-one-gigabyte WARC files, resulting in over five terabytes of transferred materials. Transferred content comprised one year's worth of daily collections of web sites linked from crawled syndication feeds and a sampling of YouTube video materials linked from those sites. The entire process -- from content selection to testing and initiating transfer through to completion -- took less than two weeks, most of which comprised data transfer time, which was not optimized through any particular means. Even though transfer of all materials took the bulk of the time, it is notable that the distributed availability of the third-party platform enabled us to begin testing the received materials as soon as the first bag transfer was completed. Because of this, we could begin to analyze transformation tools as well as intermediate and target formats, and to develop a conversion pipeline to run within AWS on all the received content.

Content Transformation for Scholarly Use

Although the WARC file format for web archives is an international standard widely used within the web archiving community and supported by the International Internet Preservation Consortium, it is not a ubiquitous format outside that community. There are a few reliable free/open source software toolkits for working with WARC files, but each requires familiarity with the WARC specification to get the most out of collected data. As such, although the WARC format is not a barrier to scholarly use of web archive materials per se, it nonetheless presents a non-trivial learning curve requiring a certain level of technical savvy that might turn away some potential users.

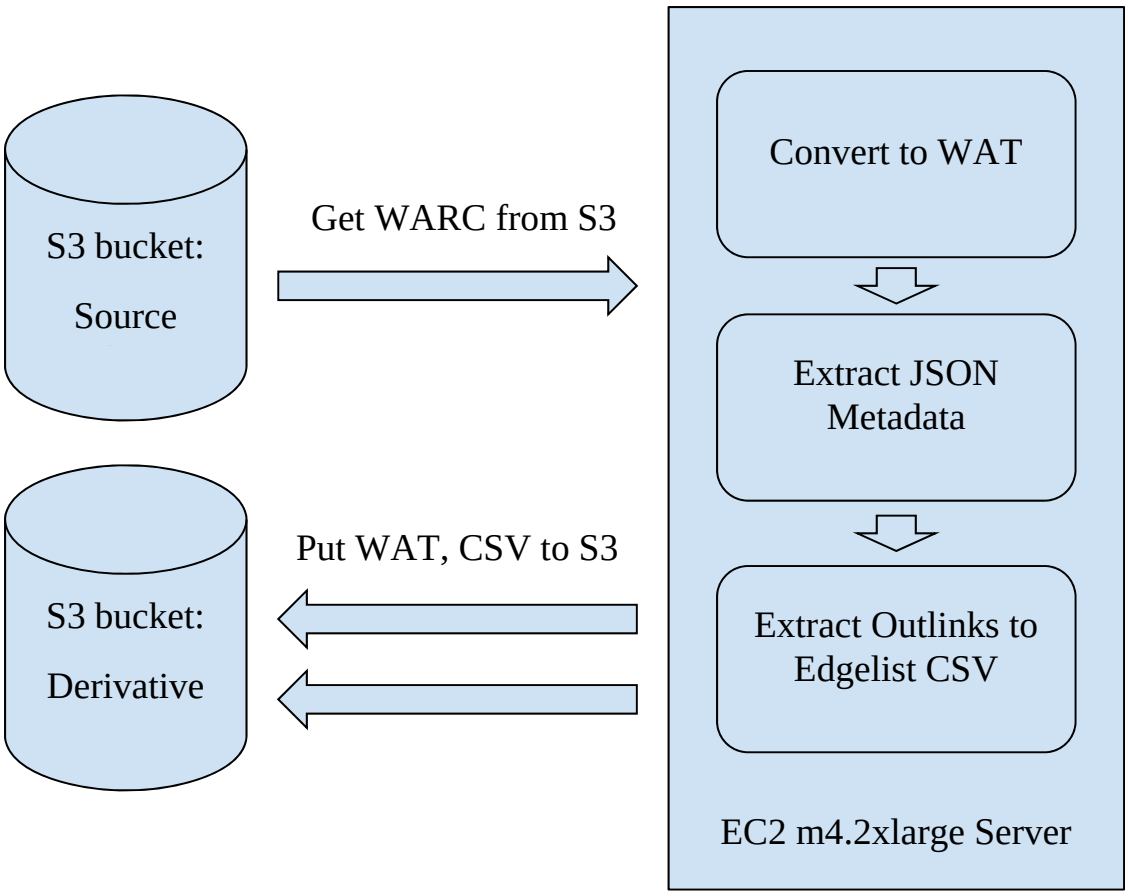
With this in mind, we considered options for processing WARC files that might lend themselves well to broader use. We opted to focus on “edgelists.” They are widely used in social science, humanities, computer science, and other disciplines to understand the structure of networks. An edgelist file consists of a list of nodes and relationships between

them (or their “edges”). This can be presented as a simple CSV file with at least two columns (typically, "source" and "target"), with more optional columns to represent edge weights or other attributes. This format is ideal for scholarship because it is amenable to graph-aware tools like Gephi, NetworkX, or any programming language, and it also works well with more widely known tools like Excel or Unix command line programs.

In the context of web archives, the source and target of an edgelist graph are typically the web page being collected (the source) and external links to web pages on other sites (the target), also known as "outlinks." These can be found within collected web content inside of WARC files, but few social scientists could be expected to extract an outlink edgelist from WARC files on their own. If the Library were to make the link contents of web archives available to scholars in an edgelist format, however, a far wider range of scholars might be empowered to dive in. Indeed, we can cite the example of the Internet Archive and their Archive-It Research Services "Longitudinal Graph Analysis" format[8] as an expression of this same concept -- a concise format detailing which pages within an archive of web pages link to which other pages -- as proving to be a useful format easy to understand for internet researchers from a variety of disciplines.

With this in mind, to complete the conceptual "delivery to a humanities scholar in a usable format,” we constructed an automated file conversion pipeline to extract edgelists from the transferred bags of Library of Congress WARC files. This pipeline, developed using the Luigi batch processing library for Python, consists of five steps, beginning with WARC files, transferring through intermediate formats, and resulting in storing simple edgelist CSV files back to S3 for later processing. This workflow is represented visually in Figure 1.

[8] "LGA Overview and Technical Details." Web. 8 December 2016. https://webarchive.jira.com/wiki/display/ARS/LGA+Overview+and+Technical+Details



Luigi (Python) format conversion workflow

Figure 1.

We were able to run this pipeline over multiple bags in parallel, processing each of the five steps listed above for each of the WARC files in each of the bags one at a time, taking on average of less than two minutes to complete the process for a single WARC file. In this way we were able to generate edgelists for each WARC file, each containing hundreds of thousands of edges, for several bags per day. Notably, this was faster than the original transfer rate. A more robust delivery workflow could automate and trigger these conversion steps upon receipt, resulting in delivery of both raw and transformed formats within essentially the same time span.

There are several benefits to this conversion step, akin to the benefits of other master/derivative format relationships. The derivative edgelist files are both more compact: edgelist CSV files, when extracted from the source WARC files in this pilot and then compressed, tend to come in at roughly 1/2000th the size of the source WARC files. In practical terms, most of the WARC files in this set are nearly one gigabyte, and most of the derivative compressed edgelist CSV files are approximately 500 kilobytes. These smaller files are much easier to transfer and store: moving only edgelist files would cut the transfer process down to minutes or hours. Even sharing the WAT intermediate format files (which

contain far more information than just edgelists, should scholars seek more detail about the contents of pages) would likely cut transfer times down by a factor of three or four, with storage overhead shrinking by a similar factor. Another advantage to these smaller, simpler edgelist files in particular is their ready suitability for batch processing in a variety of environments, which we discuss in more detail in the next section.

Options for Distributed Computing

The final step in this process would be for a scholar to perform link analysis using graph analysis tools, to study several metrics resulting from this analysis, and to produce charts and perhaps a network visualization of trends present in the network. Many tools for this work can work well on a laptop or desktop computer and handle tens and hundreds of thousands of edges quite efficiently. In the case of our selected materials, however, we dealing with a much larger set of links. A sampling of edgelist files from the pilot WARC bags indicates an average of roughly 200,000 edges (outlinks) per WARC source file, and with 300 or more WARC files in 17 bags, we are looking at a set of roughly one billion (1,000,000,000) edges in total. Scholars cannot typically study data volumes as large as one billion links using tools like Gephi or Excel on individual personal computers.

This is a case where "big data" tools like distributed computing with Hadoop make sense. The AWS Elastic MapReduce (EMR) service is ideal for such tasks. EMR may be used to preconfigure a handful, dozens, or even hundreds of EC2 virtual computer instances to work together in a single cluster, reading and writing files from S3, and optimizing computing performance using a variety of settings and optional add-on packages. For this pilot, we experimented with a variety of approaches to using EMR, such as running Apache Spark interactively using Zeppelin notebooks through the web, and using mrjob (https://pythonhosted.org/mrjob/) directly on Hadoop. For this particular project, Spark (and Zeppelin notebooks) proved to be a good choice, and we were able to write relatively simple code operating on a 20-node EMR cluster to count, sort, and compare link patterns in a sampling of transferred bags. In particular, we found 241,588,095 distinct links from crawled sites to external pages within this sampling, representing 3,886,070 unique links, and we found that by far, the major sites crawled link most to material within their own content delivery networks. With the 20-node cluster, these computations took approximately five minutes to complete. This kind of computation is not possible today using Wayback-mediated access to web archives, and Library of Congress staff themselves do not presently have the capacity to perform these sorts of analyses. As a proxy for a genuine scholar, then, we were able to demonstrate a complete workflow for preparing materials for analysis and the first steps toward application of research methods using scaled-up computing resources to begin to survey and explore data extracted from terabytes of source materials.

It is also worth noting that a scholar with little programming background is unlikely to succeed without substantial help if he or she needs to learn a unique data format like WARC, how to convert thousands of files in that format into alternate formats requiring additional investigation, how to submit jobs to a computing cluster

LIMIT OBSTACLES AND PROMOTE SCHOLARLY SUCCESS BY PROVIDING INTERMEDIATE FORMATS

requiring custom configuration and a unique set of APIs for processing, and how to negotiate the seams between related services like AWS IAM, S3, EC2, and EMR. The more we can limit those obstacles, the better the chances are for scholarly success. If staff at the Library of Congress Lab were to identify appropriate intermediate formats, aid the Web Archiving team and software developers in designing conversion workflows, assist both curators and the scholar in delivery, and offer EMR training and sample analysis code to the scholars, it is rather easy to imagine use of digital collections through channels like these increasing greatly.

Cost Considerations

The effective bill from AWS for this technical pilot was approximately $350. Ninety percent of that total was the cost of storing more than five terabytes of files in AWS S3, with the remainder the cost of EC2 virtual servers, used during the derivative generation process and during analysis with the EMR cluster. Following the logic of our earlier discussion, there are multiple ways to minimize or at least balance costs in a scenario with real scholars pursuing actual research.

The use of derivative files can result most quickly in the biggest savings. When appropriate, task-suited derivative files should be the primary unit of transfer and object of computational study, especially when scholarly methods are suited to a compact format like edgelist CSV files, as in our pilot. Ideally, such derivatives could be created at ingest or per request by the Library, rather than by scholars, to shrink the overhead of further interactions starting with transfer and verification. Although the edgelist example is perhaps an extreme one, the history of derivative file generation for access systems at LC provides a ready list of formats to consider. The new argument here, perhaps, is that there is meaningful middle ground between bulk access to raw-format masters and one-item-at-a-time web-based access system browsability of common format derivatives: bulk access to common format derivatives. These derivatives may be easily generated, transferred, stored, and meaningfully computed by scholars who are not themselves programmers, at modest costs for both the scholars and the Library.

A second major consideration relates to the balance of ownership of the third-party platform accounts. In the case of our technical pilot, we stood in as a proxy for the Library in generating derivatives, the cost of completing which might be much less when using on-site computing. At the end of the workflow, a scholar receives edgelist files, and although they number in thousands, their cumulative size is inexpensive to store. It might be easiest to think of the Library as holding the account, which a scholar would gain access to for limited use, but a scholar presenting an S3 location and secure key pair to the Library and therefore assuming the associated interim costs should be a case worth considering as well. A variety of subtle variations may occur on a given platform like AWS, where a Library-held account might cover the costs of storage in S3, while a scholar's AWS account accumulates the cost of an EMR cluster computing on data—with restricted read access—held in the Library's S3 location. Given the combination of cost- and rights-related issues around this model, it might be ideal if the Library were to develop a standard and simple set of policies regarding direct transfer to a preferred limited set of trusted third-party computing platforms. Platforms like

Amazon's AWS and its competitors accommodate these cost allocation and rights issues by allowing organizations like the Library of Congress to find patterns and define policies that fit their needs.

Third, the combined interactions between a scholar, a curator, a collection, and a multi-faceted computing environment may themselves grow costly due to the overhead of coordination and additional skilled staff needed to complement support for digital scholarship on collections as data. It seems likely even given this limited pilot that a variety of patterns may emerge where staff expertise and experience may focus development of training, workflow automation, and best practices for some categories of scholarly inquiry on particular software tools, file formats, and computing platforms. A Lab should plan for this kind of iterative development. With expertise from a variety of departments, and through thoughtful coordination and planning, a set of patterns supporting particular combinations of research methods, digital collection formats, and computing workflows will emerge over time. A successful Lab would identify and host development of these new workflows. The Lab staff would work closely with peers throughout the Library of Congress to improve, streamline, and manage the refinement and deployment of the workflows.

Implications for Lab Design and Goals

We can extract several observations from the pilot relevant to the design of a Lab focused on increased access to digital collections at the Library of Congress. The active engagement of the Web Archive Team throughout the process represents collections staff through the Library; all want to see their collections used, and in particular, they want to see the time and effort invested in digital collections bear fruit in new discoveries and scholarship over time. At the same time, to ask busy staff to add new workflows and usage patterns to their already extensive task lists might be too much. A Lab staffed with professionals with a combined range of experience stewarding the full lifecycle of digital collections, designing and deploying automated workflows for format conversion and material delivery, managing distributed computing platforms and developing training and outreach materials highlighting their potential, and navigating the communication styles and patterns between Library users, reference librarians, systems specialists, and curatorial staff could instead take point on coordinating all of this new work.

Another area of activity a Lab might lead is the development of guidelines for connecting digital collections with differential access permissions to Library users with a range of projected uses, working closely with legal counsel and with curatorial staff again to understand the affordances and requirements of third-party platforms, as well as the effects of scale and complexity in workflows on rights and potential uses. A closely related activity might require financial analysis of cost patterns in sharing the burden of platform services with scholars, and comparing various scenarios against the total ownership cost of on-premise equipment designed to support similar patterns of work.

We must also consider the need for outreach and training. Services and digital collections made newly available must be packaged and promoted to reach their intended audiences, and visiting scholars, students, and citizens alike will need training suited to their particular needs and conducive to new computational methods. In concert with this external focus,

training and outreach need to be concentrated internally, offered to and engaging Library of Congress staff in a variety of roles in understanding what might be possible, the work required to make it happen, and how staff might take on manageable portions of responsibility in seeing new developments through to regular functions of existing departments, even if that process is likely to take several years at minimum.

Outreach and Engagement with Scholars: Observations from the User Community

An important part of designing a service is to engage with potential users to determine their needs and how to most effectively answer those needs. We sought to engage with existing researchers at the Library of Congress, academic departments looking for digital resources with which to engage their students, and scholars, researchers, and other attendees from the Library of Congress Collections as Data event in September 2016. The observations and insights from these communities helped us shape our vision for the Library of Congress Lab.

To confirm user needs and expectations for a Lab, we held a focus group with incoming Kluge scholars as well as a one-on-one interview with an established Kluge scholar. These discussions were intended to identify the support a targeted user-base would be looking for in a Lab at the Library of Congress. We asked the Kluge scholars to share with us the tools and analytics they use in relation to digital content, what roadblocks they may have encountered in accessing digital content at the Library of Congress, and their overall experiences in using digital content in their scholarship. In both the group and individual setting, there was great interest in a Lab and excitement about how it might help them transform their scholarship.

A recurring theme in conversations with Kluge scholars was that they felt ill-equipped to use Library of Congress digital collections in their studies and were uncertain where to find the help they needed to use the digital collections effectively in their research. In many cases, the Kluge scholars reported not knowing which tools to use, what kinds of digital scholarship techniques might best be employed in their areas of study, and even a lack of familiarity with what the various kinds of digital scholarship techniques might be. One scholar commented: "It's particularly hard not knowing what the right tools are to do various kinds of work -- the world of tools is vast. How does one know the best place to start? It's also challenging to know what is possible in applying digital scholarship practices to my field of research."[9] This sentiment is in keeping with observations of the role of a digital scholarship center in the reports from the Center for Networked Information and was confirmed as well in our discussions with representatives of current digital scholarship centers as is reflected later in this report.

KLUGE SCHOLARS HAVE A DEMONSTRATED NEED AND DESIRE FOR SUPPORT IN DIGITAL SCHOLARSHIP

[9] Conversation with Kevin Schwartz, Kluge Scholar, 28 September 2016

The Kluge fellows expressed interest in learning how to manage the content they are creating as part of their research. Many of the scholars are taking photos of materials at the Library of Congress and are not sure how best to organize and store these files. They also expressed interest in sharing their files with the Library so that others could benefit from the images they have created. Most of the scholars expressed confusion about how to request digitization services at the Library of Congress. While this might not fall under the purview of a Lab or its services, it is an example of the kind of query that a Lab should be able to redirect to the appropriate staff. The Library of Congress Lab needs to be able to assist interested users in getting in touch with the right parties. To do so, the Lab staff will need to have a clear understanding of the various roles, responsibilities, and services available to users throughout the Library. It is one more example of why a Lab would need to be fully integrated into the Library, aware of and connected to all the services the Library of Congress offers: to connect users with the right divisions and appropriate staff to assist them in their needs.

Many of the challenges expressed in these discussions illustrated that the scholars desired and would benefit from an introduction to digital scholarship techniques and tools. For example, the Kluge scholars mentioned that they would like to do text processing but are not sure how to proceed. They would like to explore corpus linguistics but do not know how to put together a corpus. They are aware of interesting work being done related to geographic analysis of names but they are unsure how to do it or what the best tools might be to use. They would like instruction on how to connect text and census data with geospatial data layers. They would like assistance in how to think of textual materials as data or through visual paradigms. They would like to have a better way to place limitations on initial searches as well as to perform visual searches with limitations. The Kluge scholars all reported having a sense of what can be done in digital scholarship but not a real clarity about the full span of possibilities. They would like a Lab to provide a sampling of possibilities and provide them a sense of what digital scholarship can do, what tools are available and good for what applications, and with whom they might be able to work. They are looking for clear direction in how to develop the skills they might need to apply digital scholarship to their research. They are especially interested in being exposed to interesting or provocative projects and the software used to create them. One scholar commented, “I’d like help in determining how I can use digital tools to move my research forward and what kinds of scholarship can help me in doing that. I’m willing to learn, I’m willing to teach myself tools and build my own databases, but I would value some guidance and direction.”[10]

To deepen our understanding of the ways in which a Library of Congress Lab could serve the user community, we also conducted a phone interview with Miriam Posner, coordinator of the Digital Humanities Program at UCLA. We conducted this interview to deepen our understanding of potential relationships between a Lab environment and academic education in digital scholarship. In our discussion, Dr. Posner made it clear that for universities and colleges educating students in digital scholarship, freely available digital materials for instruction sessions are important, but even more important is the context and framing they are given. In teaching digital humanities and digital scholarship techniques, it is important to have data sets that are too large to work with by hand -- thereby forcing the

[10] Conversation with Kevin Schwartz, Kluge Scholar, 28 September 2016.

students to learn to adjust and manipulate the data programmatically -- but not so large that students cannot come to grow familiar with the data over the course of a class project. It is also valuable to have data sets with detailed metadata, ideally expressed in common formats like CSV or spreadsheets. Most valuable of all is to have access to time with a subject matter expert who knows about the digital collection and can speak with the students about what the students are finding through their work with the data.[11]

These two categories of user needs - the broad survey of tools and techniques needed by advanced researchers such as the Kluge fellows, and the focused packages of datasets with considerable yet manageable size and descriptive metadata suitable for beginner and intermediate digital methods practitioners like those at UCLA - are both indicative of a need a Lab can fill well, and in a manner that dovetails neatly. A survey of techniques and tools can be designed to feature selections from digital collections that meet the needs of both groups: an approachable scale still requiring new technical skill, and a focus on available materials that may be supported by curatorial expertise. We can imagine Lab staff working with curators, catalogers, conversion specialists, and educators to design a series of learning collections that support these needs well.

This educational function for the Lab can also address a direct statement of concern we heard at the Collections as Data event, referenced earlier in this report, that those visiting Library of Congress collections for the first time -- even experienced researchers -- can be overwhelmed by the volume and variety of materials available. An initiative like defining entry-level and intermediate digital collections for new users and for researchers building and refining their personal methods and skills can put a "human-sized" interface on digital collections as data. The critical function the Lab can play here, then, is to ensure there is an interface to the Library of Congress and its digital collections that everyone can relate to -- from young students to advanced scholars, including materials highlighting traditionally underrepresented groups, making practical connections with people who might not usually have such access. The Lab can help ensure visitors can both see themselves in the data and get started working with the data. The sizable digital collections and deeply knowledgeable curatorial and reference staff at the Library of Congress enable a special ability to partner with academic institutions working to educate students in digital scholarship and from all walks of life.

Insights From Other Labs:
Observations From the Field on the Role of the Lab

We performed a series of interviews intended to give us insight into the efforts already underway at various organizations both in the United States and internationally. In our interviews, we found examples of digital scholarship centers, research services, and labs organized around training in methods and tools for students and faculty researchers, enabling development of innovative services to staff within an organization, and driving

[11] Phone Interview with Miriam Posner, UCLA Digital Humanities Program Coordinator, 4 October 2016.

transformational approaches to digital collections access and processing workflows. Each of these arrangements fits the needs and priorities of its respective host institutions.

We found examples of different kinds of institutions making different choices along these lines and finding differing levels of success relative to the core missions of their organizations and the objectives of their Labs. And we find that allowing for the possibility of failure and empowering staff to fail productively can be a defining measure of Labs -- using such struggles and their outcomes as input to future decisions and changes to workflows, staffing, training, and service offerings

Common themes emerged from our interviews with leaders in digital scholarship. A comment we heard repeatedly in our conversations was that digital scholarship centers, scholars' labs, and other computing and scholarship support work is being added to the work of libraries. In several conversations, we also learned that identifying the lab space or scholarly space as a temporary measure was a strategy to create substantive support for and engagement with lab efforts. The implied expectation was that the lab environment needed to be a unique space at its inception; over time, however, successful lab-developed services would be integrated into the suite of library services offered by the larger institution. It was suggested that one of the goals of the digital scholarship center or the scholars' lab was to effectively make itself obsolete. For example:

> "Digital Scholarship Centers are being added to the work of libraries. In order to take steps into this space in a substantive way, many organizations need to deliberately identify and define what these centers offer. While some of what digital scholarship centers do will eventually become a broader library service, defining it separately now helps to create attention and clarity which is needed by library users, staff, and administrators."[12]

In another conversation, we heard a similar assessment:

> "The goal should really be for 3-5 years of a well-planned space in which the Lab can develop. This time should be used to wedge the door open and should be a way to measure progress toward the home institution's broader focus. There needs to be public accountability for what a Lab space is, does, and how it is reintegrated back into the institution."[13]

In other interviews, we heard that while the role of the Lab was conceived of as transformational, it was intended to be more permanent. At the Rijksmuseum, there is strong interest in forming and maintaining Research Services as a dedicated but separate department to support object-based research. They are not planning to reintegrate that service into general services of the museum at a later date. This is partly because the services and users at the Rijksmuseum are different than those of a university library. The Rijksmuseum Research Services primarily serves curators, researchers, internal staff, external researchers, and students. One of the roles of the Research Services unit is to help

[12] Phone interview with Dale Askey, McMaster University Library, 6 September 2016.

[13] Phone interview with Ben O'Steen, British Library Labs, 4 October 2016.

these user groups become aware of their research and scholarship needs and how to go about answering them.

> "We can't just give our users what they want because they don't always know what is possible or what they need. It is like the quote that was attributed to Henry Ford, 'If I had asked people what they wanted, they would have said faster horses.' We are helping our users transform their scholarship. That transformative quality means that we need to help them be aware of the possibilities that are out there to make their research more effective."[14]

HELP RESEARCHERS BE AWARE OF POSSIBILIITES

Services provided by the space differed in the various institutions, but a common theme across each was supporting the significant number of users needing assistance in deepening their understanding of how digital tools and analysis can assist them in their scholarship. At McMaster, the Digital Scholarship Center offers the full spectrum of support "from researchers just beginning to do digital scholarship to those who are already expert. We find that the majority of researchers need very basic assistance. They are becoming acquainted with how digital scholarship can deepen their research and add to their domain expertise."[15] This assessment is also in keeping with our discussion with the Library of Congress Kluge scholars about their needs and interests as mentioned above. The idea that the Lab should help users transform their work is a recurring theme and one that is reflected in the kinds of services offered at many different organizations.

British Library Labs provides data in bulk as one of its main services. This is partially in response to how digital scholarship is changing research and what users are trying to get out of research. "People aren't looking for one thing anymore. Researchers are no longer sure of what item they are trying to discover. They are looking for aided exploration, a discovery that is the research process, and the ability to analyze usage and words and more."[16]

AIDED EXPLORATION, A DISCOVER THAT IS THE RESEARCH

Researchers are also looking to engage with collections in mediated ways that allow them to do research in aggregate that would not be allowed given copyright restrictions with individual texts. Indiana University and University of Illinois Urbana-Champaign co-host the HathiTrust Research Center (HTRC), which is supported via membership fees and grants. The HathiTrust Research Center provides deliberately non-consumptive[17]

---

[14] Phone interview with Saskia Scheltjens, Rijksmuseum, 23 September 2016.

[15] Phone interview with Dale Askey, McMaster University Library, 6 September 2016.

[16] Phone interview with Ben O'Steen, British Library Labs, 4 October 2016.

[17] "'Non-consumptive research' is the term digital humanities scholars use to describe the large-scale analysis of a texts—say topic modeling millions of books or data-mining tens of thousands of court cases. In non-consumptive research, a text is not read by a scholar so much as it is processed by a machine."

The Poetics of Non-Consumptive Use. http://www.samplereality.com/2013/05/22/the-poetics-of-non-consumptive-reading/. Web, 9 December 2016.

computation on a digital corpus without regard to copyright status. The HTRC high-performance computing center allows individuals and groups to use the corpus and create subsets for targeted analysis. The HTRC Data Capsule virtual machine environment allows researchers to run their own algorithms without downloading material in bulk. Previously, the HTRC Portal allowed subsetting and the use of a small set of text mining algorithms limited to processing public domain content. Given the Second Circuit Court ruling that supported write/read/transform non-consumptive research as fair use under current copyright law, the full HathiTrust corpus of digital content is now being piloted. In addition, HTRC also supports the distribution of subset datasets to user groups from public domain materials. They require that the users comply with a few sets of restrictions. HathiTrust sees a steady number of requests for this service, but there are rarely more than two or three requests in the queue at any one time.[18]

Another user service explored by the British Library Labs is providing focused, temporary search engines. They will spin up search interfaces with an index for just one collection. This allows users to target their discovery work to collections or materials that are likely to have an impact on their work. They also offer targeted search indexes for multiple, thematically linked collections, providing a reintegration of content. This is a significant benefit to researchers and scholars who need to be able to discover related content, but might otherwise be limited to searching materials based on provenance rather than theme.

Because different organizations offer a variety of different services to meet the needs of their audiences, the staffing and technical needs from service center to service center and lab to lab differ accordingly. However, we repeatedly heard during the interviews that the organizations find significant value in having staff with a wide variety of digital scholarship skills. In many cases, the lab environment not only allows experienced users to employ technology, the Lab also introduces the work of digital scholarship and the ways in which it can be performed. Dale Askey identified that at McMaster University, the majority of their Digital Scholarship Center users were looking for exposure and introduction into the field. "Therefore, we don't need deep staff expertise in specialized areas of digital scholarship in order to serve our researchers; we need staff with broad digital scholarship knowledge who are good at teaching and explaining."[19]

Another theme that emerged from these conversations was the relationship between the Lab and curators. Curators and reference staff know what users want to do and the directions they want to explore within the collections. We heard often that a Lab requires strong ties to the reference and curatorial staff of an organization. A Lab conceived of as a service center rather than a research body within the organization can succeed only if it can support the work of the researchers, scholars, and other users interested in the digital collections of the organization. These users would all benefit greatly from engagement with the curatorial and reference staff of the Library of Congress as well as the technical and methodological assistance that the Lab might offer.

[18] Phone interview with Mike Furlough, HathiTrust, 7 October 2016.

[19] Phone interview with Dale Askey, McMaster University Library, 6 September 2016.

During each interview we asked each person to identify what some of the keys to success had been for the Labs in their organization. We noted a wide range of answers, many of which could be relevant to the Library of Congress:

- "Our DSC success emerges from having four key elements: 1) staff assigned to the Centre (not split with other departments) with broad digital scholarship experience who are good at teaching and explaining; 2) dedicated space; 3) money -- specifically a discretionary fund that allows the center to respond quickly to interesting opportunities is key; 4) the ability to control technology. It's important to have a technology stack under the Centre's control, or AWS or Rackspace. Of course, for this last factor, it is critical to have someone administering it who understands how to support research IT--which differs from the rules of a production environment--as well as continually learns about new technology that can support research."[20]

- "Find the pain points of the institution. Find the pain points for the researchers. There are your significant opportunities to transform."[21]

- "Be careful not to limit a Lab to self-identified researchers or developers. People who call themselves artists, explorers, etc. will be hesitant to come into a space that is labeled for someone else. Terminology will be a gatekeeper even if that is not intended. Leave your terminology as open-ended as possible and you'll get more users."[22] This recommendation aligns with some of the comments made by Bergis Jules, Marisa Parham, and others at the Library of Congress Collections as Data event. Guarding access maintains the status quo. A Lab should be organizationally transformative, assist in transforming the research of those who use the digital collections, and be open to all users and for all uses.

- "Try to develop pathfinding projects -- projects that perform the library's mission and also invent the steps between where you are starting and where you want to be. Name the projects, name the space in charge of them so you can point to the projects and successes. There need to be quantitative measures of success."[23]

Overall, the recommendations for organizations looking to start a Lab emphasize the importance of a well-trained, versatile staff willing to help users at every level of expertise but especially willing to help users new to digital scholarship find their way. They focus on defining the work of the Lab in relation to the organization as a whole. They emphasize the importance of making a Lab a space for all and allowing it to be a transformative space in which risk-taking is acceptable. In addition to many recommendations, we heard a strong warning as well:

---

[20] Phone interview with Dale Askey, McMaster University Library, 6 September 2016.

[21] Phone interview with Ben O'Steen, British Library Labs, 4 October 2016.

[22] Phone interview with Ben O'Steen, British Library Labs, 4 October 2016.

[23] Phone interview with Ben O'Steen, British Library Labs, 4 October 2016.

"There is one pitfall to avoid: Digital Scholarship Centers don't need fancy 'bling' technology for which there is no demonstrated and documented need; these are often a waste of electricity and funds. Digital Scholarship Centers should invest their money in technology that their researchers will use or into developing their space or building staff numbers rather than on something that is visually arresting or exciting just for its own sake."[24]

[24] Phone interview with Dale Askey, McMaster University Library, 6 September 2016.

RECOMMENDATIONS

The Lab offers the Library of Congress an ability to build a community around providing access to and developing insights from digital collections. The recommendations in this section advise how the Lab should be conceived and formed, articulate a possible vision and values for the Lab, identify some of the steps necessary to implement the Lab, and make suggestions about the phasing and timeline of the work.

A Foundation for Success

These recommendations are foundational to the success of the Lab. It is necessary to address these issues to ensure the long-term function of the Lab at the Library of Congress. The lab must support the entire Library, be closely connected to the entire Library to provide a seamless service to users, increase access to the digital collections, and seek to engage as many users from as many user communities as possible.

SUPPORT LIBRARY OF CONGRESS GOALS

- It is critical that the Lab be structured to support the Library of Congress's organizational objectives and goals. The goals of the Lab must be aligned with and support the Library of Congress strategic plan. The Lab must be seen throughout the Library of Congress as providing unique and useful services to scholars, researchers, and other users in connection to their digital scholarship needs. The Lab must also be seen as connecting users to existing expertise and services in the Library of Congress's various reading rooms, divisions, and departments.

SUPPORT ALL LIBRARY OF CONGRESS DIGITAL COLLECTIONS

- It is imperative that the Lab staff have a comprehensive sense of services offered throughout the Library of Congress and how to help users navigate departments appropriately to make use of the services they need. The Lab cannot function as a stand-alone department. It must be able to support work on all Library of Congress digital collections that are accessible to scholars. The Lab must be able to work with curatorial and reference staff to support users. Users should be able to get the same support regardless of whether they engage first with the Lab or the reading rooms.

INCREASE ACCESS

- A main goal and metric of success for the Lab needs to be increased access to the Library of Congress digital collections. This is a measurable standard of success that can be quantified and will help demonstrate one of the many values of having a Lab. The Library of Congress should establish the methods by which measuring use of the digital collections makes the most sense and how use changes over time, especially after formation of the Lab.

ENGAGE USERS

- The Lab needs to engage with users across boundaries. Conceived of as a transformational space, the Lab should seek to support all users interested in working with digital collections. While it is important to serve the traditional scholar and researcher, the Lab should work to serve others interested in using Library of Congress digital collections whether in unexpected scholarship topics, in support of

communities not typically engaged with the Library of Congress, or other users and uses. Supporting the widest possible user base will help the Lab in achieving its goal to increase access to Library of Congress digital collections. It will also help the Lab in its goal to be transformational by encouraging connection between unexpected user communities. Breaking down traditional expectations and widening opportunities for the greatest variety of users possible should be a primary goal of the Lab.

Addressing the Logistical and Technical Challenges

The Library of Congress houses substantial infrastructure enabling digital collections ingest, lifecycle stewardship, and public access to materials through well-defined, easy-to-use, web-standard interfaces. Its cumulative investment in staff, network and computing resources, community building around workflows and standards, and in defining and fulfilling requirements for digital preservation at the national and global scale as well as empowering individual citizens to archive their materials personally should all be lauded. The commissioning of this report stands aside those accomplishments, asking in earnest, "How can we make digital collections even more available?" By design, then, when considering how the Library of Congress can structure a Lab to experiment with answers to that question, we must consider both sides of the equation: how can Library staff make more of its materials available in formats and at scales amenable to new forms of scholarship, and what can scholars do now with technology that necessarily changes what they can and will ask of the Library?

The latter question is perhaps easiest to address. We have continued to see dramatic increases in computing power and decreases in barriers to access and cost. After an extended period during which cultural heritage institutions worked to make their holdings available via the web, we find that scholarly methods have continued to evolve based on shifts in what's possible. Just ten years ago, a tenured professor might have required substantial grant funding to build a computing cluster to sort through millions of images or petabytes of text efficiently, and an independent researcher might have sought an affiliation with an institution just to gain access to expensive software such as geospatial information systems. Today, high school students can have a powerful, open source GIS toolkit on an inexpensive laptop, and graduate students can command hosted, pay-as-you-go computing clusters for pennies on the hour through a competitive marketplace of cloud computing services. A scholar studying color patterns in thousands of mid-century images or extracting topics from millions of pages of historical newspapers using machine learning techniques is no longer limited by computing power or access to software. They are limited, instead, by access to the materials themselves. A handful of early examples of bulk access to collections as data such as imagery from the Rijksmuseum, metadata from the Cooper-Hewitt, and the diverse holdings of the Internet Archive led to a number of creative applications, but these remain exceptional cases rather than examples of common practice, although we are encouraged by

seeing more releases such as the recent HTRC announcement of a text feature set extracted by thirteen million HathiTrust volumes.[25]

For researchers to take scholarly advantage of high-end desktop tools and cluster computing service they require access to digital collections at scale. For researchers who do not have access to their own at-scale computing environments, they need for that scale to be manageable through a low-cost, high-throughput commercial service such as those offered from Amazon, Google, or Microsoft. Our technical pilot demonstrated this model. With the ready availability of Amazon Web Services, S3, EC2, IAM, and EMR tools, we modeled transfer, receipt, transformation, extraction, and analysis of hundreds of millions of links observed within terabytes of web archives content. In some ways, the transformation, extraction, and analysis of these links were the easy part -- with a basic understanding of AWS affordances, and with the digital collection at hand, the rest of the work is similar to any other scholarly methodology, although, of course, we did not subject our pilot process to any form of rigorous review required of serious research. Understanding how to process data through AWS can be learned, and the scholarly methods applied therein will continue to be refined, but the pilot taught us as much about what a Lab might create to cross the chasm of getting digital collections materials to scholars using third-party platforms like AWS:

- A standard approach to establishing formal trust between Library staff, scholars, and third-party services such as AWS -- in the case of the pilot, a secure key exchange providing access to a secure storage location created solely for this project -- must be defined and made easy to implement;

- Patterns for transferring materials into third-party hosted services must be established and standardized through workflow software development and establishment of appropriate record-keeping systems;

- Derivative formats amenable both to bulk transfer and to scholarly inquiry must be defined and documented for major digital collection types, whether from among existing standards and specifications or through new experimentation, and software tools to automate these transformations within delivery workflows must also be developed and tested.

Each of these areas of development requires close collaboration among scholars, curators, reference and technical staff, and legal counsel. Each will require trial and error, with mistakes made along the way, and each will need refinement over time as experience shapes approaches and as scholarly use of computing resources continues to evolve. A Lab could be an ideal place to start to test creative solutions and to define requirements for further development based on their results.

These challenges are driven by the nascent ability of scholars to incorporate large-scale computing and even sophisticated desktop-level methods into their processes. Returning to

[25] "HTRC releases new dataset with features extracted from over 13 million volumes." Web. 8 December 2016. http://ischool.illinois.edu/articles/2016/12/htrc-releases-new-dataset-features-extracted-over-13-million-volumes

the question of how to ensure digital collections are even more available, there are a parallel set of needs the Lab might address on changes within the Library of Congress itself:

- Library of Congress staff must be encouraged to identify, define, and implement this new middle ground between long-term storage of raw content and web-based access to derivative formats in the context of search-and-discover collection browsing applications. This work is a natural extension of previous success in implementing API-level access to Chronicling America and the Prints & Photographs Online Catalog, standards such as Z39.50, OAI-PMH, and IIIF, and enabling bulk-level access to Chronicling America newspaper data and the Linked Data Service for authorities and vocabularies;

- Library of Congress staff must also be engaged in learning about how new scholarly methods use new computational resources and must be empowered to train with and make use of the same tools in their daily work that researchers can apply as part of their scholarly methodology;

- Lab staff can serve in guiding, leading, and training roles in these activities, translating scholarly needs into opportunities for new tool and training development, and working closely across departments to define and balance costs and priorities.

Learning From the Success of Others

There were an extensive number of insights from other labs detailed in the interview section above. While all the experiences shared were of interest and valuable, the following ideas were ones that we felt to be most relevant to the Library of Congress and important to consider while developing a Lab.

1. *THE LAB IS SERVICE ORIENTED.* The main function of the Lab should be to serve users and to support Library of Congress staff in serving users. The Lab should provide instruction and education about digital scholarship projects, methods, tools and techniques. The staff of the Lab should not seek to perform their own research as a primary activity. Any research generated from the Lab by Library staff should be the product of staff fellowships (see recommendation below) or a useful byproduct of other Lab service efforts. The Lab can and should seek to help curatorial staff to integrate scholarly methods and results from studies performed on the Library of Congress digital collections back into the Library whenever they prove to be useful and relevant.

2. *THE LAB HELPS USERS TO GROW AND DEVELOP THEIR RESEARCH.* Users of the Lab will learn more about what is possible. Their research and scholarship at the Lab with Library of Congress digital collections will help them arrive at new and exciting conclusions -- perhaps in unexpected ways.

3. *THE LAB IS AN EDUCATIONAL SPACE.* It will help users new to digital scholarship get their bearings and provide support in their attempts to learn new

tools and techniques. The Kluge Scholars have demonstrated need for this service and many others will benefit from it as well.

4. *THE LAB ENABLES ORGANIZATIONAL TRANSFORMATION.* It is a laboratory in which to test and refine services and techniques for supporting Library of Congress users that can be reintegrated into the Library of Congress as a whole. The Lab will permit ground breaking service to be refined and routinized. As that happens, workflows with the Library of Congress digital collections will shift to incorporate the tools and techniques developed by the Lab. Eventually, as scholarship continues to become more and more digitally based, the services provided by the Lab will simply be part of the overall suite of services the Library of Congress offers.

Featured Recommendations

These recommendations help shape the vision for the Lab and its connections to the Library of Congress as a whole. They suggest approaches for dealing with logistical issues that may arise when forming a Lab.

- *THE LIBRARY OF CONGRESS LAB SHOULD NOT BE A DIGITAL COLLECTIONS READING ROOM.* The Lab can and should provide support for the unique needs of digital content, including services that support the use of the digital collections in ways that the reading rooms currently cannot. Special attention needs to be focused on supporting potential users wherever they may be. The digital collections themselves should remain primarily connected to the reading rooms, curatorial staff, and reference staff that support them. The Lab should offer additional services for specific activities related to use of the digital collections. While users may use the Lab as a point of entry for Library of Congress digital collections, the Lab should seek to connect users to the reference and curatorial staff for context and domain expertise whenever relevant.

- *THE LIBRARY OF CONGRESS LAB SHOULD BE A PHYSICAL SPACE IN WHICH LAB STAFF ARE CO-LOCATED.* It is important for Lab staff to confer regularly and to support one another in providing service to the user. It is also important for curatorial and reference staff at the Library of Congress to have a location to go to where they can find support for users interested in their digital collections and for their own professional development.

- *IT CANNOT BE OVERSTATED HOW IMPORTANT IT IS FOR ALL DIGITAL COLLECTIONS AT THE LIBRARY OF CONGRESS TO BE AVAILABLE TO USERS THROUGH THE LAB.* It is antithetical to the success of the Lab for some of the Library of Congress digital collections to be off limits unless there are clear rights

restrictions prohibiting use. If there are restrictions to certain collections, they must be made publicly clear. The Lab must be able to support use of as many digital collections as possible in the widest variety of ways possible. Where consumptive use is not allowed, the Lab must be able to explore mediated and aggregated use of the digital collections. The Lab must be able to explore a variety of techniques for providing access to the digital collections beyond just on-site.

- *FOR THE LAB TO SUCCEED IT MUST SERVE THE LIBRARY OF CONGRESS AS AN INSTITUTION.* The goals of the Lab should be tied specifically to the Library of Congress strategic plan as well as to the goals of Library management. The Lab should engage as many divisions as possible, support the divisions in their efforts to serve their users, reconnect users entering the Lab to the divisions whenever possible to encourage more thorough understanding of the digital collections, and refine techniques that could be reintegrated into the Library as workflows to help the divisions better understand their digital collections.

- *IT IS IMPORTANT TO THE SUCCESS OF THE LAB THAT IT BE EMPOWERED TO RESPOND QUICKLY TO CHANGES AND REQUESTS AS THEY ARISE.* It would benefit the Lab to be supported by funding not tied to the annual budget. While no-year money is a great option, it is hard to come by. The Lab could seek funds from a variety of sources to support agile responses to changing scholarly needs. Even a small fund could allow the Lab to pivot quickly in changing times. There are several funders who have supported work in digital scholarship recently. It would be well worth investigating opportunities for funding around a Library of Congress Lab.

- *THE LIBRARY OF CONGRESS SHOULD DEVELOP A STANDARD AGREEMENT* for Lab users in which they recognize the restrictions the Library has set on the digital collections and any obligations users have while working with the collections. A standard agreement will streamline access to the collections, decrease staff time needed to process and respond to the most typical requests, and increase access to the Library of Congress digital collections overall.

- *THE LAB IS AN OPPORTUNITY FOR THE LIBRARY OF CONGRESS TO ENGAGE IN NEW WAYS WITH AS MANY COMMUNITIES AND USERS AS POSSIBLE.* The Lab is an environment in which the Library can support underrepresented groups working with and on the digital collections. In envisioning intern or fellowship opportunities for scholars and developers at the Lab, the Library of Congress should consider how to bring in people from a variety of groups including Black Girls Code, Code 2040, and others. Organizations that support minority scholars should be identified and engaged.

- *FULL-TIME PROFESSIONALS WORKING IN THE LAB SHOULD POSSESS A BROAD SET OF SKILLS,* a strong service orientation and willingness to teach, a determination to find solutions on behalf of users, and should be grounded in the technical aspects of digital collections. They do not need to be curators, expert coders, or scholars themselves, although such skills would be welcome. It will be in developing collaborations with multiple experts outside of the Lab that Lab staff will find their greatest successes. It might serve the Lab well to mix some experienced staff members familiar with multiple collections and divisions at the Library of Congress with more junior staff and external hires new to the Library who are technically inclined and committed to supporting digital scholarship. Additional expertise can be added as needs require during different phases-- perhaps through temporary assignments, as described in the following section. In all cases, Lab staff should approach their work in the Lab comfortable in the knowledge that their users will have needs they cannot anticipate.

LC Lab Plan for Execution

To speak more concretely of the Lab, we have to define its location in space, its potential permanence, and its structure within the Library of Congress as an organization. Traditionally, differential formats and collection themes led to a combination of physical reading rooms holding reference and key works from collections, work spaces for scholars, meetings, and presentations, and office space for staff with holding and processing space for materials. Over the years and among the collections, these components have varied a great deal, with materials often accumulating off site, and staff space shifting frequently. Although we might imagine a physical reading room for the Lab with access to collections as data festooned with large screen visualizations and workhorse computers used as material transfer stations, we do not recommend developing such a public-oriented space as a focal point. Instead, much like the objectives of the British Library Labs and its time-limited mission, we recommend a primary focus on the Lab as something the Library of Congress needs now, and over a short- to medium-term, to be a fulcrum for identifying, defining, implementing, and mainstreaming changes in how the Library of Congress provides access to its digital collections.

We do not recommend a focus on building a public reading room for digital collections. The value gained from a public reading room requires that unique collocation of physical materials, reference works, staff, and workspace which does not exist with strictly digital collections accessible from anywhere. Similarly, we do not recommend carving out a Lab department as a permanent fixture on the organizational chart of the Library of Congress. Technological change will remain constant, and new formats will continue to arise, but the opportunity for a Lab today centers on the need to provide digital collections as data, to empower scholars to work with those collections using the tools and techniques they prefer, and to empower Library of Congress staff to define, implement, and iterate over new digital collection conversion and delivery workflows optimized for the recent wave of innovation in scholarly inquiry. It seems possible, based on the experience of British Library Labs, that these kinds of changes can be implemented within a ten-year horizon. Perhaps the clearest measure of a successful Lab at the Library of Congress would be seeing Lab staff make their work obsolete by transferring responsibilities and services to other units during its final years.

Success requires support from the highest levels of the organization, beginning with the new Librarian of Congress, as well as from chiefs and division heads of key units to be directly involved in the early work of the Lab. It seems quite likely that early success in supporting scholars and seeing scholarly methods succeed with more and more digital collections will draw more interest and support from across the institution over time, even if such support is not present at first. However, we do not recommend waiting for broad institutional support within the Library of Congress before beginning the work of a Lab. We suggest, rather, that the Library of Congress pursue targeted projects as a means to define and refine the proposed services of the Lab.

PURSUE TARGETED PROJECTS AS A MEANS TO DEFINE AND REFINE LAB SERVICES

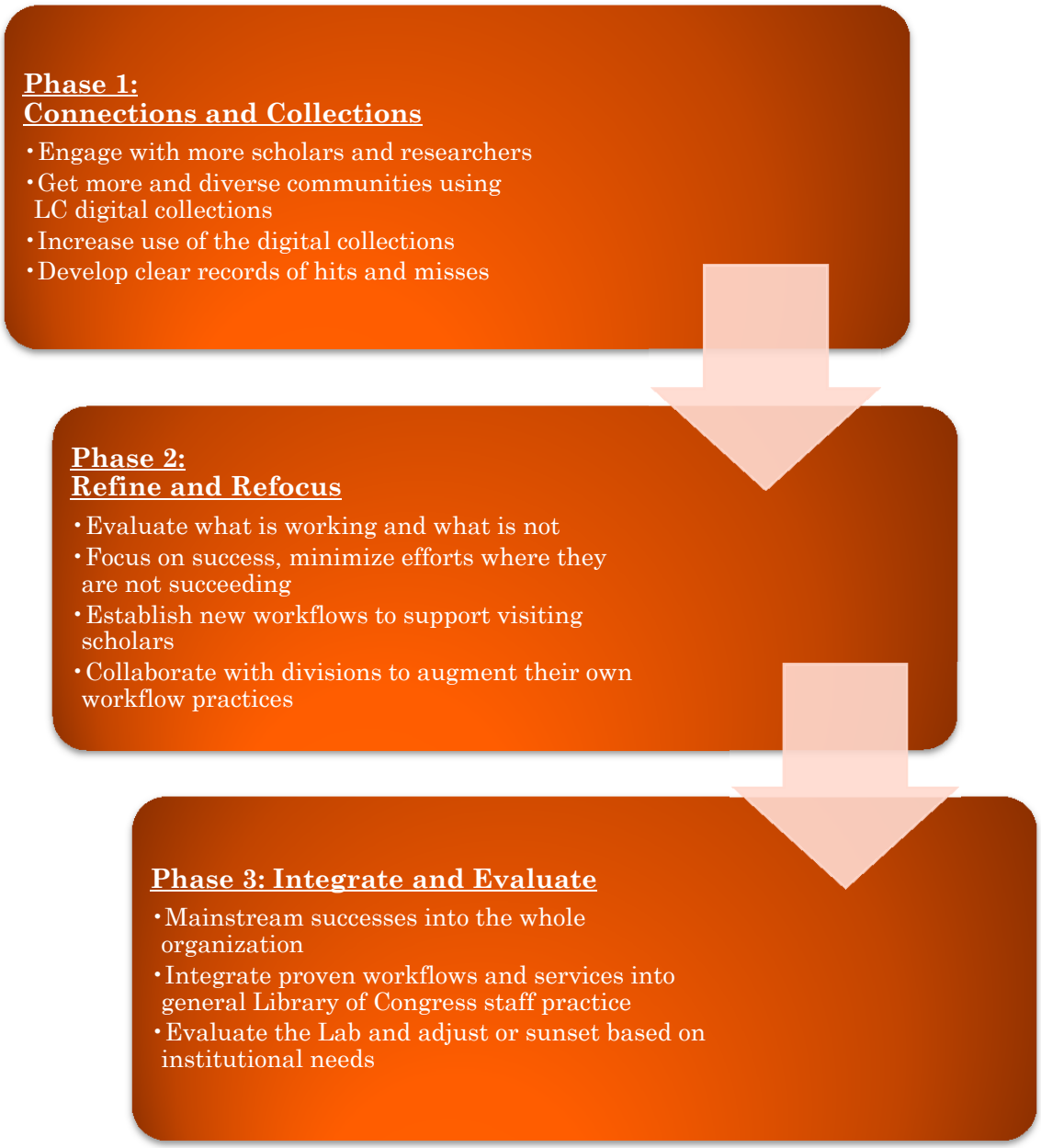


Figure 2.

Incremental Steps Forward

There are immediate and incremental steps the Library of Congress can take to support the formation of a Lab while pursuing necessary support and resources. These pathfinding projects can demonstrate the value of a Lab both internally within the Library of Congress organization as well as building interest and engagement with external users. It is important that these be identified as projects intended to help determine the path of success for a Lab. Successes from these projects can then be documented and shared to establish and grow support for developing the Lab.

Digital Scholarship Showcase

One example of an immediate step is to provide a showcase event to incoming Kluge scholars. In response to the scholars' expressed needs, the Library could model some of the training and exposure to digital scholarship a Lab would offer in a scheduled event coinciding with the arrival of Kluge scholars. This event would be an opportunity for the Library of Congress to invite digital scholarship practitioners to demonstrate some of their techniques, preferred tools, and show some of the research products that they have created. The Kluge scholars could connect with local practitioners of digital scholarship. They would learn from and become a part of that community. It would also be an opportunity for Library of Congress staff to introduce the Kluge scholars to the digital collections available for research and the current services and support available in a structured way. Library of Congress curatorial and reference staff could highlight their collections and service offerings with special emphasis on projects for which digital approaches were central to research methods. This event could be considered a prototype for the training and outreach services the Library of Congress Lab might offer.

Intern/Fellowship/Residence Model

In the first phase of British Library Labs, their work focused on inviting the winners of challenge competitions to be resident scholars at the British Library, working with Labs and other staff to see their competition-winning ideas through. The spirit of this model is present in the recent NEH-sponsored Chronicling America Data Challenge, where Library of Congress digital collections offered in bulk to contestants served as a rallying point for scholarly investigation. Artist Jer Thorp, a featured speaker at the Collections as Data event, mentioned that an artist-in-residence program might similarly forge new interdisciplinary connections while featuring digital collections and their creative use. We recommend that in the early days of the Lab at the Library of Congress, a combination of these methods be used to invite scholars, artists, students, and many other user groups to demonstrate creative and scholarly applications of Library of Congress digital collections. A mixture of competitions like the Data Challenge, open applications for residencies or visiting fellowships, and perhaps close coordination with existing programs such as the Kluge Fellowship and Kluge Fellowship in Digital Studies can be held to identify strong candidates. Winning submissions could result in funding for residencies at the Library of Congress. During these residencies, visiting fellows could work directly with Lab staff to investigate their ideas on a larger scale, using a combination of digital collections and computing resources that might drive development of new workflows, identification of intermediate file formats well-suited to investigation, and creation of training materials for students, scholars, citizens, and staff interested in pursuing similar lines of work. It might be possible to collaborate with external organizations within the government and with private foundations to fund initiatives like these. The costs of supporting residencies can be relatively small for a large potential payoff of seeing broader collection use, the training opportunities inherent in the interactions

between fellows and staff, and the benefits of cross-organizational collaboration to support such a program.

Network of Scholars

The Library of Congress's strength in convening could be particularly valuable in connecting scholars studying digital collections, faculty teaching digital methods, and curators at the Library of Congress and other cultural heritage institutions seeking to increase pedagogical outreach. As Miriam Posner of UCLA described in particular, the development and use of curricular materials centered on digital collections and research methods applying sophisticated computing tools is a rich area for potential Lab support. Dr. Posner described two particular sets of attributes that could make a project more likely to succeed in her classroom: first, a digital collection must be of large enough size that students must compute over it rather than observe it manually, and the most useful collections have useful and consistent descriptive metadata. These attributes are readily achieved using tools and techniques all of us have already. The second success factor Dr. Posner highlighted is the availability of scholars and experts in the subjects of the collection material, be they historians, archivists, or even expert curators. In her teaching, Dr. Posner invites such experts to speak with her students -- typically virtually -- to provide background and insights into the materials and their context, and she has found that her students relish the opportunity to make these connections with experts.

The Lab staff could be tasked with outreach to faculty teaching digital methods like Dr. Posner and her peers at other institutions, curators and archivists at the Library of Congress with deep knowledge of digital collections and their importance, and representatives of peer cultural heritage institutions with similar aims. The Lab could convene one or more events bringing these individuals together with the aims of exchanging ideas and curricula, building a network of experts willing and available to visit classrooms and other settings to represent collections, and identifying digital collections to target for preparation of future teaching materials. The outcomes of successful events like this could include increased use of digital collections in classrooms, a wider variety of materials made available for such use, a growing network of experts, and a clearer picture of how to present and deliver digital collections for the particular use case of teaching digital scholarship methods. In concert with these events, or in a parallel series, the Lab could convene resident scholars and competition winners as well, extending this network of digital scholars to incorporate even greater flow of innovative ideas and results. We can also imagine partnerships with HathiTrust to engage winners of the HTRC Advanced Collaborative Support Awards, with British Library Labs and winners of its annual competitions, and additional partnerships with many other organizations and researchers participating as well.

The Lab staff should also cultivate relationships with the researchers, teaching faculty, residents, and other experts who might participate in these events. These relationships will

help the Lab execute a matchmaking role for as opportunities arise, and share updates about Labs progress and scholarly results.

Opportunities for Staff Engagement

The Library of Congress Lab could be a space for Library of Congress staff to train in digital scholarship services and solutions and to begin to bring those skills into their home division. Staff could perform rotations in which they were on detail to the Lab for a period of three, six, or nine months during which time they could support the work of the Lab team as well as increase their own skills in digital scholarship. These rotations would help the Lab educate the Library of Congress staff about the services on offer. They would also help the broader staff by increasing their abilities to engage with their digital collections and directly serve their more technical users.

In addition, the Lab could also offer Library of Congress staff fellowships in which the staff of the Library could use a similar program of temporary details to perform scholarly work with Library of Congress digital collections. In these staff fellowships, staff would produce research and/or scholarship that would promote Library of Congress digital collections. Library of Congress staff experts in their subjects would have the opportunity to use the Lab, and Lab staff would support the staff fellows in learning the tools and techniques just as they would support other users. The staff fellows would get to experience the Lab as a cutting edge space in which to do influential work that matters using digital collections.

Convening the Labs Community

Repeatedly during our interviews we were reminded of the potential of the Library of Congress to act as a convener. As we observed during the Collections as Data event, when the Library invites accomplished speakers and interested guests to gather and share ideas, people will travel great distances and tune in online from all over the world to follow along and participate. CNI has hosted workshops on digital scholarship labs in 2015 and 2016. The Library of Congress could partner with CNI to continue to collaborate on next steps for this emerging community. In addition, the Library of Congress could use its significant convening power to connect historically marginalized groups with digital scholarship opportunities as well as deepening the relationships with more traditional Library of Congress connections. Hosting a conference on Labs could be an opportunity for the Library of Congress to work with leaders in the field as well as help establish new interest from broader groups. An event that sought to bring in representatives from the vendor community, different sections of the academic sphere who have challenges or tensions in their scholarship that might be aided by Lab access, as well as more creative representatives from journalism, science, startups, and even other government agencies would help to produce a more diverse and broader community as well as inspire creative solutions to common challenges.

Bulk Data Access

A Lab can be a service for Library of Congress users, staff, and to the Library as a whole. One key service of the Lab would be a complete, exhaustive website detailing all bulk data collections, APIs and services (such as IIIF, feeds, etc.), and experimental interfaces available from LC. This service would meet several goals at once: it can serve as a starting point for researchers seeking to dig into data; it can be a jumping off point for teachers and students looking to get started with methods and tools for digital inquiry; it can also be a focal point for LC staff, facilitating a shared understanding of what might be possible with digital scholarship. A complete site for access to collections as data would include sample datasets, perhaps of a scale suitable for classroom use as describe above, connected with training materials such as sample code, programming notebooks, and presentation video as appropriate.

Computing Resources for Staff Supporting Digital Scholarship

Ben O'Steen from British Library Labs emphasized the importance of flexibility in the computing environment used by Labs staff. Although security, budget, and support requirements define the standard workstation and applications available to the majority of staff at institutions like the British Library and the Library of Congress, it is important to empower staff supporting digital scholarship to with the ability to try new software, to work with multiple versions of applications on multiple operating systems, and to have access to at least some of the variety of cloud computing resources available through major platforms such as those offered by Amazon, Google, and Microsoft. It can be impossible to support digital scholarship without the ability to prepare documentation, test new techniques, and troubleshoot reported issues using a computing environment that looks at least somewhat like scholars use themselves. The same set of needs will likely require access to flexible storage pools, compute clusters, database and information retrieval services, and other computing infrastructure to support temporary projects that might be experimental, only semi-public, in support of small groups of local or remote users, or otherwise outside the typical bounds of typical federal Information Technology standards and practices. This combination of special needs might require special allowances, but cannot be overlooked as an important step in enabling staff to encourage creative uses of digital collections.

Acknowledgements

We conducted a series of interviews with leaders in the digital scholars lab, research services, and scholarly computing space. For their time and the many insights they provided, we extend our thanks to:

- *DALE ASKEY* at McMaster University Library, for discussing with us his experiences and providing insight into the elements that have made the Lewis & Ruth Sherman Centre for Digital Scholarship at McMaster a success.
- *MIKE FURLOUGH* at HathiTrust, for providing examples of some of the forms of mediated and aggregated access that are engaging to users and for encouraging broader community engagement.
- *BEN O'STEEN* at British Library Labs, for sharing with us his insights about making sure a space welcomes the widest possible number of users and the importance of developing pathfinding projects.
- *MIRIAM POSNER* at UCLA, for helping us understand the educational use case and the needs of the student and researcher more clearly.
- *SASKIA SCHELTJENS* at the Rijksmuseum, for speaking with us about the critical components of providing services to an organization and to users who are still developing their understanding of the space.

We would like to extend thanks to the John W. Kluge Center, its staff -- especially Mary Lou Reker -- and all the 2016 Kluge scholars, with special thanks to Kevin Schwartz, for their time and insights into the work they would like to be doing with Library of Congress digital collections.

We are grateful to Abbie Grotke, Grace Thomas, and the rest of the Web Archiving Team for their time and consideration in supporting our work on this project. We found the transfer of the Web Archiving Collection to be a useful pilot that illuminated several opportunities and challenges. Special thanks to David Brunton in OCIO Web Services for his invaluable assistance in helping the Web Archiving team to transfer the content.

Thanks to the American Folklife Center for their invaluable insights about how researchers are using their collections and the kinds of work that researchers would like to see supported at the Library of Congress.

Our appreciation to the NDI team for including us in the Collections As Data event and conversations. Thank you to Kate Zwaard, Abigail Potter, Jamie Mears, Michael Ashenfelder, and the entire team for providing feedback and acting as our points of contact for information concerning the Library of Congress. Special thanks to Chris Adams of Web Services for his insights and thoughtful responses.

About the Authors

Michelle Gallinger is Principal of Gallinger Consulting where she provides technological insight into decision-making processes for libraries, museums, archives, and businesses. She gives strategic planning guidance; develops policies, guidelines, and action plans for her clients; offers stakeholder facilitation services; and coordinates collaborative technological initiatives. Gallinger's clients include Harvard Library, Institute of Museums and Library Services, Metropolitan New York Library Council, Ithaka S+R, and the Council of State Archivists (CoSA). Prior to consulting, Gallinger worked at the Library of Congress developing the initial strategy for and led the creation, definition, and launch of the National Digital Stewardship Alliance in 2010. Gallinger performed policy development, strategic planning, program planning, and research and analysis at the Library of Congress. She has also designed and managed digitization services, overseen the formation of and policy development for the Colonial Williamsburg Rockefeller Library's Digital Library, and worked for Gartner Dataquest.

Michelle Gallinger holds a bachelor of arts in English literature from Reed College and a master of arts in English literature from the University of Virginia where her research focused on humanities computing. She can be reached at mgallinger@gallingerconsult.com.

Daniel Chudnov is a librarian, software developer, and data scientist with twenty years of experience designing and deploying applications that solve a wide range of information problems. He has worked at the Cushing/Whitney Medical Library and the Center for Medical Informatics at Yale University School of Medicine, the MIT Libraries, the Library of Congress, and the George Washington University Libraries. During his career he has contributed to several prominent free/open source software projects, as well as several large-scale digital library resources including World Digital Library and Chronicling America. He was principal investigator on grants from IMLS and NHPRC to develop Social Feed Manager at GWU, an application supporting services to faculty and student researchers studying social media. He has written extensively about innovative software and services in libraries, and is faculty at District Data Labs, where he coordinates research activities.

Daniel holds a bachelor of arts in Economics and master of science in Information, both from the University of Michigan, and a master of science in Business Analytics from the George Washington University School of Business, where he is also an adjunct lecturer, teaching Data Management for Analytics. He is Principal at Chudnov Consulting, an independent consultancy based in Washington, DC, where he may be reached at d@chud.co.

APPENDIX A

Web Archiving Team Input to Digital Scholars Lab report pilot

The following is a write up of the experiences of the Web Archiving team throughout this pilot. This perspective provides additional context and insight about the transfer of data and work that made it happen.

S3 transfer process

In order to transfer the RSS content from LC systems to the Amazon S3 bucket set up by Dan, David Brunton first installed the AWS Command Line Interface (CLI) on the Web Archiving Team's four virtual machines (VMs) utilized for indexing. The VMs are able to access a mounted version of the spinning-disk access copy of web archived content. Once navigated inside the correct directory, in this case, the RSS content, it was quite simple to use the "put" command with the --recursive option to copy from the LC storage to the S3 bucket, for example:

s3cmd put --recursive [RSS bag name]/ s3://[bucket name]/ [RSS bag name]/

The initial transfer succeeded, but lacked the second [RSS bag name] path specification, transferring all loose files into the bucket. To account for this, subsequent transfers included the bag name in the path, which created a directory-like structure for all content in the bucket.

For the pilot, transfer was not optimized. Since uploading multiple terabytes to an S3 bucket was an entirely new process at LC, we were careful not to overload the virtual machines to respect concurrent processes utilizing the same resources. To limit the strain, LC staff transferred one bag of content on each VM at a time. During the two week transfer time frame, there may have been hours between the completion of a transfer and kick off of the next bag (i.e. if the transfer finished at 3 a.m. the next one would not start until staff arrived to LC at 8 a.m.). As this process is streamlined, a simple script could kick off the next bag when the previous completes and more research could be done on how to optimize the processing VMs.

About the Library's Web Archive and Current Researcher Access

As of the end of FY16, Library had collected 1.06 petabytes (1067 terabytes) and over 14 billion documents. Currently there are 11,288 records in 22 described event or thematic collections available via www.loc.gov/websites for research use. Web Archive items that have been cataloged do not correspond easily to "one record per website"; rather a record may describe different or specific parts of a website that has been archived, or an entire website or organization's web properties, or a candidate who has run for office in many Elections with varying URLs that have been archived. The 22 collections represent only about a third of the Library's web archives; the rest are awaiting cataloging and other processing resources to describe them and make them publicly accessible via the Library's website.

In addition to the catalogued item records and collection descriptions, the Library has made content available via its OpenWayback up through June 2016 (we have a one-year embargo and plan to release new content this way every six months or so). Although there have been requests to include a URL search in the Project One Web Archives interface for better access, that is not currently implemented, so not many researchers are aware that additional content is available via Wayback access (http://webarchive.loc.gov/all/*/[insert URL HERE]). The Library's collections are Memento compliant and are accessible via this search http://timetravel.mementoweb.org/, which we have on occasion shared with staff and researchers in lieu of a URL search at LC.

About Deduplication

Web Archiving deduplication, where the byte and URL are exact to another copy, became possible when Heritrix 3.0 and the WARC standard became available in 2009. The Library implemented deduplication because of the space savings. For example, in test crawls, the house.gov and senate.gov sites saved 80% of space for each monthly crawl, so it would require storage space of only 1 Terabyte for every 5 terabytes of content. The WARC standard provided for a "revisit record" which documented which WARC the original content was stored in, and the Wayback replays the deduplicated content as if it were captured on that date. So for users, it is transparent that Wayback is replaying content captured months or weeks before.

The Library has implemented a policy of getting an undeduped crawl (for each crawl event) as a baseline. As content is crawled, it is either NEW or EXACTLY the same as a previous copy. So, for every crawl there is new content that is used to deduplicate against subsequent crawls. It's an iterative process.

From 2010 to 2015, that content was baselined in January each year, or when new crawl event began (for example, Election crawls have begun at various points in time). With a contract change in 2015, the contractor was instructed to perform the baseline undeduplicated crawls at the "beginning of each task" which generally was meant to be beginning of the fiscal year (October), however tasks are not always lined up with the fiscal year any longer, and some crawling in 2016-2017 has shifted those baselines to other dates entirely, for instance the RSS crawls used in this pilot were baselined in October 2015 but were not in 2016 due to the nature of those crawls.

Deduplication of our crawls in this manner may create issues in creating data sets for researchers for specific collections, as the baseline crawls would likely need to be included for the data sets to make sense when being used by researchers. For example, a researcher interested in content collected by the Library for a particular time period (say, at the end of a Congressional term), might need to receive data from a full year of archiving for the revisit links in WARC to resolve).

The Library plans to do regular and clearer documentation of baseline crawls, particularly in light of this pilot and implications for sharing data sets of web archive content.

About the Web Archive Collection Structure and Crawling Buckets

The Library has, since the beginning of the web archiving activity, collected in event and thematic collections, which have been developed by recommending officers and subject experts around the Library. This is in part because of the early pilot years and decisions about copyright and approaches to permissions – by including sites in named collections, it was easier to argue fair use and make other policy decisions. After 16 years of archiving, we still take this approach, even as permissions policies have shifted.

From 2000-early 2010 the Library initiated separate crawls each time a new collection was proposed. During most of this time, we were only doing a handful of collections per year, so managing this wasn't so bad, however setting up new collections took a few weeks to a month to establish the appropriate mechanism through our contract crawling agent. As the program matured, and there were requests to do more collections and to begin archiving new collections more quickly, we decided to shift to a "crawl bucket" mode as of around March 2010, which grouped things mostly by frequency (and sometimes by format, such as RSS feed crawling) rather than by collection. This helped the program become more nimble in terms of adding collections and seeds in and out of the crawls, saving collection start up time and costs. However, this means that as of March 2010, all the collections (the exceptions being the U.S. Election archives which are big enough to crawl on their own still, and the RSS feeds used in the pilot), are grouped into bags that are NOT collection-specific.

The implications of this for future use by researchers using a Data Scholars Lab could mean that researchers requesting access to specific event or thematic archives (besides the election content) may have to receive data that includes other collections archived at the same time, and most collections have seeds with varying frequencies. Some examples (not taking into consideration deduplication issues mentioned above):

- The Library could easily provide a researcher interested in the Visual Image Web Archive, since it was collected 04/05/2006 - 11/30/2006 and stored in separate collection –specific bags
- The Library could easily provide a researcher interested in the Election web archives data, since they are always crawled separately
- A researcher interested in all .gov materials collected by the library would be a challenge; we have collected many domains across many event and thematic collections in a variety of frequencies, in collection specific bags and also in all crawling buckets.
- A researcher interested in our Papal transition archives would easily receive the 2005 transition archive (with collection –specific bags), but the Papal Transition 2013 Web Archive was done with seeds in both the monthly and weekly crawling buckets.

The Web Archiving team has access to this information for each collection but it is not readily available in one place and would require some amount of research time depending on the specific collection requested.