IFLA WLIC 2018
KUALA LUMPUR

Submitted on: 23.04.2019

Institution as Social Media Collector: Lessons Learned from the Library of Congress

Kate Zwaard
Digital Strategy Directorate, Library of Congress , Washington, DC, USA

Moryma Aydelott
Library Services, Library of Congress, Washington, DC, USA

David Brunton
Platform Services, Design and Development, Library of Congress, Washington, DC, USA

Melissa Crawford
Digital Strategy Directorate, Library of Congress, Washington, DC, USA

Abbie Grotke
Web Archiving Team, Library Services, Library of Congress, Washington, DC, USA

Nicole Marcou
Office of the Librarian, Library of Congress, Washington, DC, USA

Jaime Mears
Digital Strategy Directorate, Library of Congress, Washington, DC, USA

Jared Nagel
Digital Strategy Directorate, Library of Congress, Washington, DC, USA

Abigail Potter
Digital Strategy Directorate, Library of Congress, Washington, DC, USA

Michelle Rago
Office of Communications, Library of Congress, Washington, DC, USA

Steve Short
Library Services, Library of Congress, Washington, DC, USA

J. Mark Sweeney
Deputy Librarian of Congress, Library of Congress, Washington, DC, USA

E-mail: lc.labs@loc.gov

Abstract:

Social media has become a valuable tool that has helped connect, facilitate, discover, and encourage use of the Library of Congress's collections and content. The Library has cultivated trust and approachability as both a collector and user of social media. As a result, the institution is more accessible than ever before.

Keywords: social media, born-digital, web archive

1 INTRODUCTION

The Library of Congress's mission is to engage, inspire, and inform Congress and the American people with a universal and enduring source of knowledge and creativity. The Library is responsible for serving the United States Congress (as well as the rest of the United States Federal Government), the scholarly community, and the general public. The Library of Congress (which we refer to in this paper as the Library) has been using social media both as a collector ***and*** a user to execute its mission. We will describe our use of social media in both of these roles beginning with the Library as a collector.

2 LIBRARY AS A COLLECTOR

2.1 Collection development policies and The Digital Collecting Plan

As of 2018, the Library had 167 million tangible items on approximately 838 miles of bookshelves. Among these collections are 24 million general collections books and extraordinary special collections, including the world's largest collection of maps, music scores, films, sound recordings, comic books and telephone directories. Each week, about 20,000 tangible items arrive at the Library through U.S. Copyright registration. Of these, subject specialists determine which of these the Library will keep, according to the Library's collection development policies (https://www.loc.gov/acq/devpol/cpsstate.html).

While these numbers describe the tangible material, the digital collection at the Library is harder to quantify. It comprises both material that has been digitized and our born digital collections, including more than 13,000 electronic serial titles, half a million electronic books, and 17 billion unique files in the web archives. And we continue to expand the digital formats we are collecting and the mechanisms we are using for collection. The Digital Collecting Plan, (https://www.loc.gov/acq/devpol/CollectingDigitalContent.pdf), published in 2017, explains this in greater detail. We are collecting for the use of researchers both now and in the future and exploring access models for contemporary access to these items.

2.2 Collecting Via Web Archiving

The web archiving program at the Library began in 2000, before the launch of contemporary social media platforms. Collecting efforts began as a pilot program to explore methods by which the Library would collect, preserve, describe and make available web archived content. The program has grown to over 120 event-based and thematic collections in various stages of production (https://www.loc.gov/programs/web-archiving/about-this-program/). As with its tangible collecting program, the Library collects web archives selectively. The process of creating a web archive collection begins with librarians. Recommending officers -- librarians with deep subject matter expertise -- recommend items for the collection. They determine the collection focus and select content for the web archives in their subject area. These recommending officers decide on a thematic and event-based collection, then select appropriate websites to add to the collection. Some examples of the Library's web archiving collections are Science Blogs; Web Cultures; Olympic Games; U.S. government sites from the Legislative, Judicial, and Executive branch agencies; foreign government sites; campaign web sites and political parties documenting U.S. and foreign elections; non-profit organizations; journalism and news; and international organizations. While most web archives are collected as a part of one or more event or thematic archives, the Library also preserves other sites within its general web archives.

2.3 Collecting Social Media via Web Archiving

While the primary focus of the Library's web archiving program is to preserve websites of organizations and individuals, there is an attempt to selectively archive the social media properties to which these sites are publishing content. For instance, when the Library archives the website of a government organization, such as the U.S. Department of State, staff use "scoping instructions" to capture related social media properties and content hosted on other domains to archive fuller representation of the web presence of the organization or person.

Sometimes social media platforms are the only places where content is being published by an organization or individual. For example, primarily in our U.S. Election web archives (https://www.loc.gov/collections/united-states-elections-web-archive/about-this-collection/), social media accounts like Facebook or Twitter increasingly serve as primary campaign sites rather than as traditional websites. To properly document elections, which is one of the Library's core collecting areas, we must try to acquire this content. Consequently, the Library's web archives now include material published on Facebook, Twitter, Pinterest, YouTube, Friendster, and other social media platforms. Social media content presents technical challenges, however; depending on the platform, harvesting tools in use sometime run into difficulty in preserving this content.

2.4 Collecting Social Media through Other Means

Web archiving is not the only mechanism for capturing social media. The Library also collects directly from the publishers themselves. For example, StoryCorps, a non-profit organization that records and preserves American oral histories, launched a social media platform a few years ago. In partnership with StoryCorps, the Library collects oral histories from that platform directly using their API. For more information, please see https://blogs.loc.gov/thesignal/2015/12/acquiring-at-digital-scale-harvesting-the-storycorps-me-collection/.

The Library has also collected directly from the photo-sharing site, Flickr, to develop thematic collections around American's folk traditions. For this project, an "opt-in" method was used, asking users to tag images they would like to share. For more information, please see https://blogs.loc.gov/folklife/2016/01/celebrate-afcs-40th-with-photos-of-mytradition/. A campaign asking users to tag images is also an example of how the Library can use social media to engage users. By far the largest collection donated from a social media organization, at least by volume of entries, is the Library's early archive of tweets from Twitter.

2.5 Twitter archive

In 2010, the Library announced an agreement with Twitter detailing the acquisition of the entire archive of public tweets from the launch of Twitter in 2006. The agreement also included a continuous acquisition of public tweets going forward as a daily delivery. As a result, the Library has a secure collection of tweet text documenting the first 12 years of this dynamic communications channel—its emergence, its applications, and its evolution. As all libraries do, the Library regularly reviews their collections practices to consider environmental shifts, diversity of topics, and other factors. The Library does not typically acquire all content published on a particular platform. Instead, items are collected selectively according to collection policies (https://loc.gov/acq/devpol/). Following that model, as with all other platforms, in 2017, the Library announced a change in collection practices which ended the ongoing collection of a comprehensive set of public tweets, choosing instead to focus on selected, intentional topics. As stated in the white paper published in December of 2017, *Update on the Twitter Archive at the Library of Congress* (https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf), "The Library will focus its efforts on preserving the Twitter collection for future generations. Throughout its history, the Library has seized opportunities to collect snapshots of unique moments in human history and preserve them for future generations. These snapshots of particular moments in history often give voice to history's silent masses: ordinary people. Without the efforts of past generations, the nation might not have a collection of oral histories in the hours after the attack on Pearl Harbor (memory.loc.gov/ammem/afcphhtml/afcphhome.html) or film footage depicting San Francisco before and after the great quake of 1906 (loc.gov/collections/san-francisco-earthquake-and-fire-1897-to-1916/about-this-collection/). The Twitter Archive may prove to be one of this generation's most significant legacies to future generations. Future generations will learn much about this rich period in our history, the information flows, and social and political forces that help define the current generation."

2.6 Lessons Learned: Scale of Collecting

Collecting across many digital platforms has been a learning experience for the Library. We learned how to plan and scale infrastructure to meet the exponential growth of given content streams. As already described, the Library collects social media selectively, as it does with all other acquisitions, based on collecting scope and policies, which are often affected by legal requirements or limitations.

However, even preserving selectively, libraries have to plan for tremendous growth of content on the web, including storage considerations and processing needs that go beyond just storing the data and making it available. For example, description at scale is a challenge.

"With over 20,000 web archives among 114 ongoing and finished collections, the scale of the Library's web archive has grown significantly, presenting compelling new challenges for description along the way. To provide access at the same rate the archive continues to expand, the Web Archiving Team , representatives from Acquisitions and Bibliographic Access , and Web Services created an innovative new cataloguing approach. The approach, known internally as the minimal-record approach, "combines the descriptive talents of cataloguing librarians with the power of Python scripting to automatically create MODS records," said Grace Thomas, Digital Collections Specialist for the Library of Congress Web Archiving Team, in a blog post on August 3, 2018 about web archiving in The Signal (https://blogs.loc.gov/thesignal/2018/08/more-web-archives-less-process/).

2.7 Lessons Learned: Researcher Needs

The Library is working to understand user needs when researching social media. In the John W. Kluge Center, several fellows have investigated how social media relates to culture, history and even language (https://blogs.loc.gov/kluge/2017/06/emoji-texting-and-social-media-how-do-they-impact-language/). As one example, an Assistant Professor in a School for Communication at a major university came to the Library of Congress as a Digital Studies Fellow in residence at the Kluge

Center, researching how people have used different media technologies to buy and sell things in informal markets. The researcher worked with the Library's Chronicling America collection of historic newspapers to identify instances of fraudulent ads and bartering, and to compare these to contemporary ads on Craigslist. While the millions of historic newspaper pages are amenable to full-text search, there is currently no way to analyze classified ads as a subset of the data/text. A Library software developer suggested a crowdsourced approach for quickly identifying and tagging ads, but regardless of approach, the data wrangling necessary to separate the ads from the larger corpus could not be completed during the researcher's fellowship term. Recognizing the need to use Python to analyze the data once it is procured, the researcher took a Software Carpentry (https://software-carpentry.org/) class at the Library for a hands-on introduction and planned to create the data set after the fellowship.

In another example, the Library hosted the annual web archive datathon Archives Unleashed 2.0 in 2016, joining the Archives Unleashed team of seasoned researchers and instructors to better understand how to run future events and support computational research with web archives and social media. Over two days, teams worked with social media content from the Library's election web archives, George Washington University's Social Media Feed Manager and on-the-fly sets they grabbed themselves from Twitter to produce insightful discoveries by the 48-hour deadline. The majority of researchers had experience with data-research methodologies such as text and network analysis and were familiar with a variety of analytical querying tools. Even with this background, participants required technical and subject knowledge support throughout the process, from feedback on initial proposals to technical support with tools and unruly data sets and frequent interventions from subject matter experts to help decipher results (https://blogs.loc.gov/thesignal/2016/07/co-hosting-a-datathon-at-the-library-of-congress/).

In these case studies and the handful of others that have taken place in the past five years, it is clear the need and amount of technical and subject support to use social media data sets varies based on the nature of a user's inquiry, technical knowledge, and the constraints of available data sets. Though the Library does have deeply knowledgeable subject experts to support these researchers and help decode their results, there are no dedicated digital scholarship staff to assist in these efforts. In the meantime, the Library is working to make data sets such as its web archives (https://labs.loc.gov/experiments/webarchive-datasets/) more accessible, and continues to acquire current content.

2.8 Economic value of social media content

Selling access to historic content has become an important part of the business models for many social media platforms. This has implications for libraries managing, securing, and serving large datasets of this material, which have more commercial value now than they may have had at the point of acquisition. Social media companies making profits on repackaging or providing access to their data are making investments in discovery tools and computing power beyond the scale of library budgets. Additionally, these organizations have knowledge of and access to proprietary systems. These incentives can change the ecosystem for knowledge sharing and cooperation between non-profit archives and for-profit social media organizations for immediate access to the data.

Still, there is value for libraries in continuing the rich collecting tradition in archives of collecting material to which immediate access may not be provided, due to donor privacy or other concerns, to ensure access in the future. Researchers in the future studying our contemporary society will find great value in the social media collections, during which our search and discovery tools will mature as will the computing power we can draw from. Additionally, the access points that libraries and archives afford often have stability beyond a typical lifespan of commercial platforms.

2.9 Lessons Learned: User Privacy & Content fixity

There are many factors to take into account when considering acquiring social media. For instance, how willing are you as an institution to allow the content to be changed, and by whom? It is vital, as a cultural heritage organization, that we maintain content integrity and authenticity. With many social media platforms, content is constantly being created, edited, revised, and deleted by the original authors. As cultural heritage institutions that serve future generations as well as the current generation, libraries face many unique challenges managing this dynamic content. One should consider the legal and ethical obligations libraries and others preserving social media content have to the original authors of the content.

Each library has a role in ensuring that is content served is authentic with an established a chain of custody between ingest and delivery. A key tool the Library uses is fixity – when we acquire a digital file, we calculate a cryptographic hash on the file (sometimes referred to as a "checksum"). Ideally, this value is calculated by the depositor (e.g., the donor or vendor) and calculated again upon arrival at the Library to ensure that we have received what the depositor intended to deliver. In the case of acquisitions from StoryCorps, Flickr, and Twitter, the Library was able to establish a chain of custody in this manner.

In one recent case, having the same content in multiple archives provided evidence against claims that the Internet Archive was "hacked." Some preservation tools have built in features that support fixity. For example, one of the unique features of the tool Social Feed Manager is that they record their interactions with social media APIs in WARC files (https://gwu-libraries.github.io/sfm-ui/).

3.0 LIBRARY AS A USER

3.1 The Library of Congress uses social media to connect

Our vision is that all Americans are connected to the Library, as stated in the Library's strategic plan: https://www.loc.gov/strategic-plan/. Social media has allowed the Library to extend that connection into discovery and use of the Library's vast online collections and engage audiences in ways that are more approachable and immediate. The Library's many social media channels can be found at https://www.loc.gov/connect/ and includes Library-written blogs, Facebook pages, Twitter accounts, Instagram, Flickr, and Pinterest, as well as our latest addition, Medium. We also have video channels and podcasts.

While individual offices are empowered to craft their own social media presences, the practice at the Library is centrally coordinated. Through that coordination, the Library builds a community of practice and offers training on regulatory and legal guidelines that are particular to our environment. For example, the Library faces somewhat unusual challenges as a government agency, which are prohibited from using official channels for certain kinds of speech, such as endorsing commercial products. The Library of Congress Communications Office, Office of General Counsel, and social media specialists work together to craft policies, offer guidance, and provide training. They also host mechanisms like reoccurring brown-bag lunches for Library staff to coordinate and share best practices.

3.2 Telling Personal Stories

In 2016, The Librarian of Congress, Dr. Carla Hayden launched her twitter account. Her anecdotes about the collections provide regular invitations for people to connect the Library's holdings with their everyday lives, memories, and identities.

Social media also presents opportunities to strengthen connections within communities. Programs at the Library will hold open chats on Twitter using hashtags like #ChronAmParty or #edChat to engage with their users and to encourage conversation in a community of practice.

At the Library, we see that social media allows us to connect with new and unexpected audiences when we are amplified by ambassadors of particular communities, connecting us across disciplinary, language, and cultural differences. For example, The Library tweets a series called "Today in History," a bite-sized view into their collections. The most retweeted Library of Congress tweet of all time was from one of these tweets. It read, "Today in History: Aaron Burr shoots Alexander Hamilton in a duel in Weehawken, N.J., 1804" It was retweeted by Lin-Manuel Miranda, the author of the popular musical *Hamilton*. By the next morning the tweet had received almost 700,000 views and almost 5,000 retweets, thanks in large part to Miranda's retweet.

3.3 Beyond Words

The importance of social media lies in establishing mechanisms for two-way communication with users. The Library recognizes that information exchange is how we establish relationships and deepen connections with one another.

One way the Library is exploring two way communications with their users is by inviting their contributions through crowdsourcing, capturing the enthusiasm of the public and considering how to support their discovery with content shared from the Library.

The Library has been using crowdsourcing for more than ten years through the photo-sharing website, Flickr, to enhance understanding of their Prints and Photographs collections. On Flickr, users can comment and tag photos in the Library's collection and interact with curators, giving details and context from their own experiences, which are sometimes included in the Library's metadata about the item.

Beyond Words is an experimental project developed by the Library's Labs team to test out new crowdsourcing methods and applications. This application represents a different, complementary type of interaction -- one that is less rich, but can support more scale.

In Beyond Words, users are invited to help identify illustrations and provide captions in WWI-era newspapers. After users identify those illustrations, they are added to a searchable public domain image gallery. These images include historic maps, photographs, and political cartoons. The ability to search and reuse these presents opportunities for scholarship and creative play. Additionally, users are encouraged to share these materials through other social media platforms with their own communities, broadening the reach of the Library. All data created is released into the public domain and available immediately in a gallery view and in JSON, a format that can be used by software developers.

Since the launch of the project, the Library has seen 146,443 contributions by users resulting in 1683 marked and captioned images. They average of about 9,000 contributions a month, which demonstrates the public's interest in interacting with primary source material and in supporting libraries. The Library used this pilot as a learning opportunity, and incorporated many of its lessons into a new crowdsourcing program, *By the People.*

3.4 Crowd.loc.gov

In October 2018, the Library launched a new crowdsourcing platform, *By the People*, available at https://crowd.loc.gov/. This project invites users to transcribe and tag images of text from the Library's collections. These community-developed transcriptions will increase the usability and findability of the Library's digitized historic texts, like letters between poets or diaries of historic

figures, by making them keyword searchable and usable by screen readers for the first time.

Crowd.loc.gov is both a consumer and producer of potential social media content. As a producer, the Library uses its social media to dialog with the community and drive engagement. As a consumer, users can use their own social media presences to engage with the Library material, each other, and their own communities.

The great thing about this is that the Library can connect, facilitate discover, and encourage use all in one place. At its best, social media allows the Library to expand conversation with the public and broaden their understanding of the world through first-hand experience with the Library's collections and content. This kind of bi-directional exchange and appearing in places that users are already familiar with allows us to cultivate trust and approachability.

4 CONCLUSION

As part of fulling its mission to engage, inspire, and inform Congress and the American people with a universal and enduring source of knowledge and creativity, the Library is working diligently to ensure that the information within its collections is accurate and accessible. This is done taking into consideration available resources, various ways of capturing data, identifying tools that can aid in collection, and ensuring quality control. The Library may be the oldest federal cultural institution in the nation, but we strive to open new horizons for our users and the library community by exploring and sharing our experiences collecting and digitizing pertinent information that people around the world need.