DIGITAL PRESERVATION: THE TWITTER ARCHIVES AND NDIIPP

Laura E. Campbell
The Library of Congress
Office of Strategic Initiatives
101 Independence Ave, SE
Washington DC 20540

Beth Dulabahn
The Library of Congress

ABSTRACT

On April 14, 2010, the Library of Congress and Twitter made the joint announcement that the Library would receive a gift of the archive of all public tweets shared through the service since its inception in 2006. The media and community response was tremendous, raising many questions about how the Library would be stewarding and providing access to the collection. There are many issues to consider, from the technical mechanisms of transfer to the Library and the ongoing updates to the archive, to curatorial policies, to planning for a new type of research access to a Library collection. The Twitter archive joins a number of born-digital collections at the Library. This year the National Digital Information Infrastructure and Preservation Program (NDIIPP) celebrates a decade of digital preservation actions and discovery, working with a network of over 170 partners resulting in over 200 terabytes of all types of important digital content. Collectively, we have built lasting relationships, helped facilitate natural networks within the overall NDIIPP network, and tested new tools and services that support the work of the partners in the network. We will rely on the collective wisdom of our partners as we grapple with the challenges of curating and serving a digital collection as rich and diverse as the Twitter archive.

1 The Twitter Archives Acquisition

Twitter is a microblogging service that enables users to send and receive messages of up to 140 characters, called "tweets." Users share their tweets and others follow tweets in a social network environment. Worldwide, Twitter processes more than 50 million tweets per day and this number is growing exponentially.

On April 14, 2010, the Library of Congress and Twitter made the joint announcement [8,10] that the Library would receive a gift of the archive of all public tweets shared through the service since its inception in 2006. The media and community response to the announcement -- simultaneously tweeted and blogged -- was tremendous, beginning a very public conversation about what the Library would receive and how the Library would be stewarding and providing access to the collection.

Although to many it seems an incongruous acquisition for the Library, the Library holds a wide range of materials in many formats, and collects groups of items as well as individual items. With the receipt of the Twitter archive, the Library continues its long tradition of collecting and preserving personal stories, such as the "man on the street" interviews after Pearl Harbor; personal letters and diaries collected for the Veterans History Project; and conversations between family members preserved in StoryCorps[1]. Twitter forms part of the historical record of communication in the twenty-first century, capturing news reports, events, and social trends. Minute-by-minute headlines from major news sources such as Reuters, The Wall Street Journal and The New York Times are pushed to Twitter. At the same time, it serves as a platform for citizen journalism with many significant events being first reported by eyewitnesses. It is frequently cited as an important unfiltered record of important events such as the 2008 U.S. presidential election or the "Green Revolution" in Iran.

The Library also has a long history of enriching its collections through donations. The Twitter archive is a gift from Twitter; the agreement is openly available online.[2] After the Twitter announcement, Greg Pass, Twitter's vice president of engineering, said: "We are pleased and proud to make this collection available for the benefit of the American people. I am very grateful that Dr. Billington and the Library recognize the value of this information. It is something new, but it tells an amazing story that needs to be remembered." [8]

For its potential value, the size is relatively small - approximately 5 terabytes for all public tweets from 2006 to early 2010. This makes it considerably smaller than the Library's other web archives[3], which comprise more than 170 terabytes of web sites, including legal blogs, topical and event archives,

[1] StoryCorps is available at http://storycorps.org.

[2] The agreement between the Library of Congress and Twitter is available at: http://blogs.loc.gov/loc/files/2010/04/LOC-Twitter.pdf

[3] The Library of Congress Web Archives are available at: http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html.

election campaigns for national office, and websites of Members of Congress.

2 The Archive as Case Study

In addition to providing access to the archive, the Library also sees this acquisition as an opportunity for expanding external collaborations with digital preservation partners. The Library has worked with many of its partners to develop and test mechanisms for content transfer, and does not anticipate any significant problems in the actual transfer of the archive from Twitter to the Library. In terms of curation, however, the Twitter archive pushes the limits of traditional models of curation. The contents of the archive cross virtually all of the subject areas within the Library, making it difficult to assign any division sole custodial responsibility. The Library's research and education partners in this effort - scholars, historians, librarians, archivists and scientists - provide unique perspectives when thinking creatively about digital curatorial responsibility, user access, and Library support and services.

Dr. James Billington, the Librarian of Congress, agrees that the benefit is not only to the American people, but to the Library's goal to gain experience in best practices and procedures in support of its collections. Dr. Billington said: "The Library looks at this as an opportunity to add new kinds of information without subtracting from our responsibility to manage our overall collection. Working with the Twitter archive will also help the Library extend its capability to provide stewardship for very large sets of born-digital materials." [8] The Twitter archive will serve as a case study for the management and preservation of a large corpus of digital data.

3 Privacy issues

Twitter is donating an archive of what it determines to be public. Alexander Macgillivray, Twitter's general counsel, as quoted in the *New York Times* [9], said, "From the beginning, Twitter has been a public and open service." Twitter's privacy policy states: "Our services are primarily designed to help you share information with the world. Most of the information you provide to us is information you are asking us to make public." Under the Twitter terms of service, users give Twitter the right to archive tweets.

There will be at least a six-month window between the original date of a tweet and its date of availability for research use. Private account information and deleted tweets will not be part of the archive. Linked information such as pictures and websites is not part of the archive, and the Library has no plans to collect the linked sites. Moreover, the Library does not expect to make the collection available online in its entirety. Use will be limited to non-commercial private study, scholarship, or research.

The Library understands there are concerns about privacy issues and is sensitive to those concerns. The Library has a long history of respecting sensitive information related to its research collections and will be mindful of those concerns as it develops its access plans. Periodic public communications about the archive will set expectations for privacy and access.

4 Research Use and Access

The collections will be made available to non-commercial researchers and to Library staff for internal use. Details about researcher access policies are being developed with input from curators across the Library, taking into account principles that have been applied to existing collections. While wanting to stay consistent with the philosophy that has governed the use of its physical collections, the Library recognizes that a digital collection such as the Twitter archive presents a number of new challenges, and is exploring how to accommodate not only a wide range of research uses, but also a geographically diverse set of users.

Viewed in the aggregate, the Twitter collection can be a resource for current and future researchers to study and understand contemporary culture and political, economic and social trends and topics. The Library has assembled a set of use cases from scholarly publications, news publications and blogs, social scientists, and librarians. The use cases drive the creation of archival policy requirements related to search, access, privacy, and preservation.

For historians, Twitter provides direct witness accounts of events in real time. It also serves as a virtual timeline of communications about events, people and places. This provides an enormous amount of raw unmediated primary source material for historical research.

Daniel J. Cohen, an associate professor of history at George Mason University and co-author of a 2006 book, "Digital History", as quoted in the *New York Times* [9], said that "Twitter is tens of millions of active users. There is no archive with tens of millions of diaries".

The Twitter archive could be used to study empirically how individuals reacted to a particular historical event. If Twitter existed on on September 11, 2001, the American people would have a real-time chronicle of what people were thinking and feeling on that day. Today, it is a popular environment for "first-on-the-scene" news reports. In February 2008, the earthquake in the United Kingdom was reported on Twitter at least 35 minutes before it was reported in the mainstream press. Later that year, Twitter was used by eyewitnesses of the Mumbai terror attacks:

"Hospital update. Shots still being fired. Also, Metro cinema next door," twittered Mumbaiattack;
"Mumbai terrorists are asking hotel reception for room #s of American citizens and holding them hostage on one floor", twittered Dupreee.

Twitter can be compared to earlier sources of personal information such as diaries and letters. Many of the earlier sources contain mundane or trivial pieces of information that, in aggregate, can tell a detailed and authentic story about everyday life that is difficult to find elsewhere. David Ferriero, the Archivist of the United States, points out that historians often find value, sometimes unanticipated, in what others may see as mundane details of our lives and what they might say about our culture [2]. Paul Freedman, a professor of history at Yale University, agrees. Freedman was quoted in *Slate* as saying: "Historians are interested in ordinary life." [1] Only time will tell its value. It could be that Twitter content may be studied by future scholars in the same way that the graffiti of Pompeii is being studied by current scholars.

Social scientists are similarly interested in using the collection to study trends and patterns, such as social networks which are of interest to a wide range of disciplines from anthropology to political science, management science, sociology and communications. A 2010 study of Twitter use by social science researchers revealed that Twitter was ranked in the top three services used by researchers to spread information. [6] Researchers are increasingly interested in studying scientific networks and the spread of information inside and outside the scientific community.

Researchers may also study communities that drum up support using Twitter. In June of 2009, political dissidents in Iran used Twitter to voice opposition while Iranian newspapers were heavily censored.[4] Future researchers may choose to extract these archived tweets, using time and location data, in order to draw conclusions about the opinions and attitudes of the period.

Researchers and research organizations are excited at the prospect of exploring the Twitter's rich and varied data in the aggregate, especially with newer data mining and social graph analysis tools that can reveal trends[7]. It is always a challenge to predict how the archive might be interrogated, so user feedback and requests will be collected in the coming months and years to help the Library investigate how its can potentially expose its collections as data. The Library may also enter into technical partnerships with external agencies and organizations to develop search and visualization tools for use with the archive. The Library is currently involved in such a partnership with Stanford University, called the Computational Approaches to Digital Stewardship Project[4], which is focused on new tools for the discovery of digital collections.

5 Managing the Archive

While the primary focus of the Twitter gift is the retrospective archive - tweets from 2006 to 2010 - the Library and Twitter are working on a mutually-agreeable form for incremental updates. The terms of the gift agreement specify that the Library not make available any tweet less than six months old. Therefore, the technical committee responsible for the ongoing archive must develop a framework for receipt, ingest and management that considers this six month hold.

There are interesting issues to be addressed in the areas of receipt , ingest and management going forward:

- The number of tweets per day has been increasing significantly. This means that each incremental update will be significantly larger as we move forward.
- The terms of the gift require that the Library not make available any tweet that is less than six months old. This means that the incremental update receipt and ingest process needs to take this into account.
- The terms of the gift include only public tweets. This means that the incremental update receipt and ingest process needs to take this into account.
- What practices will be followed to verify that the data received by the Library is the same as the data sent by Twitter?
- What practices will be followed to ingest and store the large number of tweets? Will those practices be affected by the method of update, or by changes in the fields or format of future tweets?
- What processing will the Library be performing for management of the tweets or to make them available to researchers?
- What kinds of services will the Library offer to researchers and how will those affect the management of the tweets?

The Library expects to identify and analyze options that would address these issues. The Library may explore a multi-stage process for receipt that optimizes the processing flows at each stage. A multi-stage approach would also allow for each stage to be configured and tuned for best resource use and flexibility for changes in the volume and/or data. For example, the Library could establish a local (or remote) isolated staging area for receipt of update files or streams. One or more processes could then perform any required verification, processing and adding public tweets older

[4] http://cads.stanford.edu/.

than six months to the data set available for researcher use. The Library expects to explore and test technical options that could make up feasible implementation solutions.

The Library is looking forward to expanding its capabilities to take in, make available, and preserve creative content for current and future generations. The Twitter gift provides an opportunity to make progress that we hope will also benefit other institutions and partners that are addressing some of these same issues.

6 The National Digital Information Infrastructure and Preservation Program

The Twitter archive is just one of many born-digital collections that the Library has brought under its stewardship, part of a long history of working with digital content. In 2010 the National Digital Information Infrastructure and Preservation Program (NDIIPP) is celebrating ten years of digital preservation actions and discovery working with a network of over 170 partners resulting in over 200 terabytes of all types of important digital content. Collectively, we have built lasting relationships, helped facilitate natural networks within the overall NDIIPP network, and tested new tools and services that support the work of the partners in the network. The collaboration today has federal and state government partners, commercial content partners, service providers, library and archival institutional partners and international partners. We proudly consider the work of the last decade a true collaboration that reflects enormous transformation in the way libraries will work in the future.

A new outgrowth of NDIIPP is the National Digital Stewardship Alliance (NDSA), a collaborative effort among government agencies, educational institutions, non-profit organizations, and business entities to preserve a distributed national digital collection for the benefit of citizens now and in the future. The NDSA is an inclusive organization that will focus on shared work toward common community goals.

We plan to draw on our partners for assistance in technical approaches to supporting a digital archive of this size, richness, and complexity.

REFERENCES

[1] Beam, Christopher. "How future historians will use the Twitter archives." *Slate*. Web, April 20, 2010. July 11, 2010. <http://www.slate.com/id/2251429>

[2] Ferriero, David. "Tweets: What we might learn from mundane details." *AOTUS: Collector in Chief.* Web. April 16, 2010. July 11, 2010. <http://blogs.archives.gov/aotus/?p=172>.

[3] Margot Gerritsen– private correspondence to Laura Campbell.

[4] Grossman, Lev. "Iran Protests: Twitter, the Medium of the Movement." *Time Magazine*. June 17, 2009. Web. July 11, 2010. <http://www.time.com/time/world/article/0,8599,1905125,00.html>

[5] Haddadi, Meeyoung Cha Juan Antonio Navarro P´erez Hamed. *Flash Floods and Ripples: The Spread of Media Content through the Blogosphere*. Association for the Advancement of Artificial Intelligence, 2009.

[6] Letierce, Julie; Passant, Alexandre; Decker, Stefan Breslin, John G. *Understanding how Twitter is used to spread scientific messages.* Web Science Conference 2010, April 26-27, 2010 Raleigh NC.

[7] McLemee, Scott. "The Mood is the Message." *Inside Higher Ed*. June 30, 2010. Web. July 11, 2010. <http://www.insidehighered.com/views/mclemee/mclemee296>

[8] Raymond, Matt. "Twitter Donates Entire Tweet Archive to Library of Congress." Library of Congress, April 15, 2010. Web. July 11,2010. <http://www.loc.gov/today/pr/2010/10-081.html>

[9] Stross, Randall. "When History is Compiled 140 Characters at a Time." *The New York Times:* April 30, 2010. <http://www.nytimes.com/2010/05/02/business/02digi.htm>

[10] Twitter. "Tweet Preservation." Twitter, April 15, 2010. Web. July 11,2010. <http://blog.twitter.com/2010/04/tweet-preservation.html>

[11] Wasserman, Stanley; Galaskiewicz ,Joseph. *Advances in social network analysis: research in the social and behavioral sciences.* Sage Publishing, 1994.