January 2013

Update on the Twitter Archive
At the Library of Congress

In April, 2010, the Library of Congress and Twitter signed an agreement providing the Library the public tweets from the company's inception through the date of the agreement, an archive of tweets from 2006 through April, 2010. Additionally, the Library and Twitter agreed that Twitter would provide all public tweets on an ongoing basis under the same terms. The Library's first objectives were to acquire and preserve the 2006-10 archive; to establish a secure, sustainable process for receiving and preserving a daily, ongoing stream of tweets through the present day; and to create a structure for organizing the entire archive by date. This month, all those objectives will be completed. To date, the Library has an archive of approximately 170 billion tweets.

The Library's focus now is on confronting and working around the technology challenges to making the archive accessible to researchers and policymakers in a comprehensive, useful way. It is clear that technology to allow for scholarship access to large data sets is lagging behind technology for creating and distributing such data. Even the private sector has not yet implemented cost-effective commercial solutions because of the complexity and resource requirements of such a task. The Library is now pursuing partnerships with the private sector to allow some limited access capability in our reading rooms. These efforts are ongoing and a priority for the Library.

This document summarizes the Library's work to date and outlines present-day progress and challenges.

Why the Twitter Collection is Important to the Nation's Library

Twitter is a new kind of collection for the Library of Congress, but an important one to its mission of serving both Congress and the public. As society turns to social media as a primary method of communication and creative expression, social media is supplementing and in some cases supplanting letters, journals, serial publications and other sources routinely collected by research libraries.

Archiving and preserving outlets such as Twitter will enable future researchers access to a fuller picture of today's cultural norms, dialogue, trends and events to inform scholarship, the legislative process, new works of authorship, education and other purposes.

The Library of Congress Agreement with Twitter

The Library's agreement with Twitter announced April 14, 2010 provided that:

- Twitter would donate a collection consisting of all public tweets from the Twitter service from its inception to the date of the agreement, an archive of 21 billion tweets that occurred between 2006 and 2010.
- Any additional materials Twitter provides to the Library would be governed by the terms of the agreement unless both parties agree to different terms in advance of receiving such additional materials.
- The Library could make available any portion of the collection six months after it was originally posted on Twitter to "bona fide" researchers.
- A researcher must sign a "notification" prohibiting commercial use and redistribution of the collection.
- The Library cannot provide a substantial portion of the collection on its web site in a form that can be easily downloaded.

Transfer of Data to the Library

In December, 2010, Twitter named a Colorado-based company, Gnip, as the delivery agent for moving data to the Library.

Shortly thereafter, the Library and Gnip began to agree on specifications and processes for the transfer of files – "current" tweets - on an ongoing basis.

In February 2011, transfer of "current" tweets was initiated and began with tweets from December 2010.

On February 28, 2012, the Library received the 2006-2010 archive through Gnip in three compressed files totaling 2.3 terabytes. When uncompressed the files total 20 terabytes. The files contained approximately 21 billion tweets, each with more than 50 accompanying metadata fields, such as place and description.

As of December 1, 2012, the Library has received more than 150 billion additional tweets and corresponding metadata, for a total including the 2006-2010 archive of approximately 170 billion tweets totaling 133.2 terabytes for two compressed copies.

Building a Stable, Sustainable Archive

The Library's first and most fundamental activities included developing a stable and sustainable way to acquire, preserve and organize the Twitter collection.

Although the Library regularly acquires digital content, the Twitter stream is the first collection coming into the Library in a continuous stream. The Library leveraged the technical infrastructure and workflow established for other digital content in the transfer of Twitter data.

The Library runs a fully automated process for taking in these new files. Gnip, the designated delivery agent for Twitter, receives tweets in a single real-time stream from

Twitter. Gnip organizes the stream of tweets into hour-long segments and uploads these files to a secure server throughout the day for retrieval by the Library.

When a new file is available, the Library downloads the file to a temporary server space, checks the materials for completeness and transfer corruption, captures statistics about the number of tweets in each file, copies the file to tape, and deletes the file from the temporary server space.

The technical infrastructure for the Library's Twitter archive follows the same general practices for monitoring and managing other digital collection data at the Library. Tape archives are the Library's standard for preservation and long-term storage. Files are copied to two tape archives in geographically different locations as a preservation and security measure.

The volume of tweets the Library receives each day has grown from 140 million beginning in February, 2011 to nearly half a billion tweets each day as of October, 2012.

The Library is processing data from the original 2006-2010 archive and organizing the material into hourly files. This operation is necessary so the entire archive from 2006 moving forward is organized the same – by time and in hourly files. This process will be completed in January 2013.

Toward Providing Collection Research Access

As with any collection, the Twitter archive must be processed and organized in a way that makes it useable. It is not uncommon for the Library to spend months or in some cases years sorting a large acquisition to inventory, organize and catalogue the information and materials so they are accessible by researchers.

The Library has extensive expertise in managing acquisition of and access to large-volume digital collections. For example, since 2000, the Library has been collecting web sites documenting government information and policy events. Today, that archive is more than 300 terabytes in size, and represents tens of thousands of web sites. Because there was a community of cultural heritage institutions and national libraries committed to collecting web sites, standards and tools have been developed collaboratively for capturing and providing access to these materials.

The Twitter Archive represents a new type of collection. The Twitter collection is not only very large, it also is expanding daily, and at a rapidly increasing velocity. The variety of tweets is also high, considering distinctions between original tweets, re-tweets using the Twitter software, re-tweets that are manually designated as such, tweets with embedded links or pictures and other varieties.

The Library has received approximately 400 inquiries from researchers all over the world since the announcement that it would accept the Twitter archive. Some broad topics of interest expressed by researchers thus far run from patterns in the rise of citizen

journalism and interest in elected officials' communications to tracking vaccination rates and predicting stock market activity. Many inquiries would inform research with policy and regulatory usefulness, such as tracking flu pandemic, citizen responses to candidates' stances on various issues and tracking public access to court systems. The nature of the queries also varies. For example, requests range from searching for a specific hashtag term to requesting a statistically valid sample of the entire stream.

What kind of information might researchers learn from the Twitter archive?

Some examples of the types of requests the Library has received indicate how researchers might use this archive to inform future scholarship:

* A master's student is interested in understanding the role of citizens in disruptive events. The student is focusing on real-time micro-blogging of terrorist attacks. The questions focus on the timeliness and accuracy of tweets during specified events.

* A post-doctoral researcher is looking at the language used to spread information about charities' activities and solicitations via social media during and immediately following natural disasters. The questions focus on audience targets and effectiveness.

The Library has not yet provided researchers access to the archive. Currently, executing a single search of just the fixed 2006-2010 archive on the Library's systems could take 24 hours. This is an inadequate situation in which to begin offering access to researchers, as it so severely limits the number of possible searches.

The Library has assessed existing software and hardware solutions that divide and simultaneously search large data sets to reduce search time, so-called "distributed and parallel computing". To achieve a significant reduction of search time, however, would require an extensive infrastructure of hundreds if not thousands of servers. This is cost-prohibitive and impractical for a public institution.

Some private companies offer access to historic tweets but they are not the free, indexed and searchable access that would be of most value to legislative researchers and scholars.

It is clear that technology to allow for scholarship access to large data sets is not nearly as advanced as the technology for creating and distributing that data. Even the private sector has not yet implemented cost-effective commercial solutions because of the complexity and resource requirements of such a task.

Twitter chief executive Dick Costolo this year announced that the company is working on providing Twitter users with access to all their own tweets. In a July 24 *New York Times* interview he also addressed the question of a search engine enabling access to all tweets, saying, "It's two different search problems. It's a different way of architecting search, going through all tweets of all time. You can't just put three engineers on it."

In the near term, the Library is working to develop a basic level of access that can be implemented while archival access technologies catch up. The Library will consult with congressional researchers and scholars to inform this process. These efforts are ongoing and a priority for the Library. Potential scenarios include public-private partnerships and leveraging private sector investment and capacity.

Recently, senior Library officials met with Gnip senior management in Washington to explore the possibility of developing a research- and scholarship-focused interface to the archive using Gnip's existing historical Twitter product offerings.

The Library continues on a daily basis to build and preserve this important archive, with the expectation that it will be accessible to researchers on premises.

The Library is managing this collection in keeping with part of its mission to acquire, preserve and provide access to a universal collection of knowledge and the record of America's creativity for Congress and the American people. The Library looks forward to continued collaboration with the private sector and the research community as we continue to maintain and build the collection and work toward making this resource accessible for scholarship in a comprehensive, useful way.