US Library of Congress faces big data challenge for Twitter archive

The US Library of Congess is creating an archive of everything ever posted on Twitter. Researchers hope to use the tool to gain new insights into society.

"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," says the US Library of Congress's director of communications, Gayle Osterberg.
Image courtesy Clarissa Peterson, Flickr (CC BY-NC-SA 2.0).

Twitter's not just about finding out what Lady Gaga had for breakfast or sharing amusing photos of cats, you know. It's also about getting into arguments with strangers and watching sports stars make fools of themselves.

More importantly (and more seriously) though, it's an utterly huge repository of social data which researchers are just itching to get their hands on. From mapping political trends, to analyzing social interactions and examining the way memes spread across the globe, the 500-million-user-strong microblogging site is a potential treasure trove of information.

Posted on JAN 16 2013 7:54AM

Tags

big data
breakfast photos
Justin Bieber LOLcats
Lady Gaga
Library of Congress
searchability
social media
sociology
Twitter

To this end, the US Library of Congress signed a deal in April 2010 to create a complete archive of the Twitterverse. And, last week, the Library published a report outlining its progress with the project. According to the report, the library has an archive of around 170 billion tweets, stretching back to Twitter's launch in 2006. The tweets each have over 50 accompanying metadata fields and the archive now totals over 130 terabytes in size. It's expanding rapidly too, with new tweets now being added on an hourly basis via the help of social media aggregation company Gnip. This means that the archive is growing at a staggering rate of over half billion new tweets each day, up from around 'just' 140 million tweets a day in early 2011.

However, it's not storing all of this data which is a problem for the Library of Congress. Rather, making the archive easily searchable is proving tricky. A single search of the 2006-2010 archive of tweets - just one eighth the size of the entire volume - can currently take up to 24 hours. Yet, even once this searchability issue has been cracked, it still doesn't mean everyone will be able to poke around the Twitter archive. As part of the Library's agreement, it will make the archive available only to "bona fide" researchers. So far, over 400 applications have been made to access the archive, including projects looking at vaccination rates, citizen journalism, and the stock market. As of yet though, there have apparently been no applications to answer that most perplexing of Twitter mysteries: namely, how on Earth does Justin Bieber have 32 million followers?

Find out more about the big data challenges the Twitter archive poses by reading the Library of Congress's report in full here.

- Andrew Purcell

Join the conversation

Contribute

Do you have story ideas or something to contribute? Let us know!

FUNDING PARTNERS

The National Science Foundation supports the US desk under award 1242759, for sustaining and strengthening International Science Grid This Week (which recently became the Science Node).

CERN, the European Organization for Nuclear Research, supports the Science Node. The organization has played a key role in the publication since 2006, and currently hosts the European editor.

CATEGORIES

Advanced computing
Research networks
Big data
Tech trends
Community building

CONNECT WITH US

CONTACT

Science Node
Email:
editors@sciencenode.o

Website:
sciencenode.org