Twitter's increasingly influential role in journalism has prompted an accompanying upsurge in academic research, particularly around the ways in which journalists and media organizations have integrated Twitter into their norms and practices.

With 500 million tweets a day, Twitter offers researchers a potentially deep and rich stream of social media data. However, unlike historical newspaper content, which is readily available via library microfiches or databases like Lexis Nexis, much of the historical data on Twitter (what's called the Twitter firehose) is walled off in costly private archives.

Information may want to be free, but accessing and analyzing that information can be costly.

The Library of Congress signed a deal with Twitter in 2010 to build an on-site research archive but that system has still not been finalized. A progress update is expected this summer, but the archive, which now houses more than 170 billion tweets, poses major logistical challenges for the Library and the firehose reseller Gnip, which is delivering the data for Twitter. For example, a single search of the 21 billion tweets in the fixed 2006-10 archive was taking 24 hours just last year. Twitter acquired Gnip in April, prompting hopes that the archive may be operational in 2014-15, but even so, the archive will only be accessible on-site at the Library in Washington, D.C.

That source of reliable, inexpensive online access to the Twitter firehose has become almost a Holy Grail for journalism professors in the U.S. and Canada who I surveyed this June using a Google form.

Kathleen Culver, assistant professor at the University of Wisconsin-Madison, says she would like to see "portals for academics into Twitter, supported by Twitter," and an easy user interface for research. Alf Hermida, associate professor at the University of British Columbia, agrees. Hermida, who has just published a new book on social media, says such a portal could contain "shared archives of Twitter data, best practices and approaches."
Mike Reilley, online journalism instructor at DePaul University in Chicago says he wants something that will let him go deep. “I’m looking for that ‘super tool’ and am hoping someone at JS Knight Stanford comes up with one. An all-in-one tool – scrapes, archive, great search, everything. I’m tired of ‘tool-hopping’ to get work done,” he said.

**Flawed early Twitter research tools**

Early efforts such as the freemium TwapperKeeper service offered that “all-in-one” functionality, albeit with some restrictions. TwapperKeeper, which allowed users to create and download .csv files of Twitter archives, limited historical searches to 7 days earlier or 3,500 tweets (whichever came sooner). However, TwapperKeeper, which was launched in 2009, was taken over by HootSuite in 2011 and became a premium subscription product.

Some researchers then shifted to Topsy Pro, which offered trial accounts for researchers or a single annual license for $12,000, but the datasets were often incomplete.

Hermida says Topsy used its own criteria to delete tweets from the archive. “Topsy removed tweets that had been deleted from the Twitter firehose, and tweets without at least six retweets or a retweet by an influential user were removed from the search index after 30 days,” he said.

Robert Hernandez, associate professor of professional practice at USC Annenberg School for Communication & Journalism, says he has long suspected such archives could be flawed. “I always have an uneasy feeling that the archive – whether I get it formally or not – doesn’t feel as accurate or complete as one thinks,” he said.

‘Divide between data-rich and data-poor researchers’

A more robust historical search, such as analyzing the #Newtown tweets to examine journalists’ behavior in the hours immediately following the school massacre, would require subscription to a certified Twitter firehose service such as Gnip or DataSift.

These services can retrieve an unlimited amount of tweets from practically any time in Twitter’s history. But the resellers’ main focus is businesses seeking more data on how consumers view them. Thus, their pricing is aimed at institutions or corporations. The pricing plans for both companies are difficult to decipher (both sites ask users to submit a Web form for a quote) but a 2014 article put DataSift at about $3,000 a month with Gnip starting at $500 for each one-off search. Licensing fees cost an additional $0.10 per 1,000 tweets and are paid to Twitter. These licensing fees accounted for $32 million of Twitter’s earnings in the first half of 2013.

However, a $3,000-a-month subscription level or even a $500 search would be too expensive for most academics unless they were able to make arrangements...
with their institutions.

Hermida's university uses Crimson Hexagon, which charges $5,000 a year for 50 search terms or "monitors" as part of its **Social Research Grant Program**. Elizabeth Breese, senior content and digital marketing strategist, said the program seeks researchers who are "a good fit... that the research is non-commercial in nature, and that the results will be made public in some way," she said by email. The grant program provides 50 "simultaneous monitors" which can be deleted to provide for a new query, giving researchers more flexibility.

Meanwhile, a freemium Twitter scraping tool from the British company **ScraperWiki** could be the solution for cash-poor researchers. Users set a hashtag, keyword or user name as a search term and then let ScraperWiki monitor Twitter for all new occurrences. Like TwapperKeeper before it, ScraperWiki can only create new archives rather than search for historical data, but the drawbacks of the service are mitigated by the price.

Pricing starts at $9 a month for "Explorer" access to three datasets and tops out at $29 a month for "Data Scientist" access to 100 datasets. The main drawback, as discussed, is the lack of historical data. But the tool is incredibly robust and a well-planned project could return tens of thousands of rows of data for analysis and visualization as there is no maximum limit on tweets.

For example, using the Data Scientist package, a researcher could easily track the ongoing output from up to 100 users for a project such as a comparative analysis on a constructed week. In my own research, I used ScraperWiki to retrieve approximately 22,000 tweets tracing the social media development of the Tuam babies story in Ireland in May/June 2014.

ScraperWiki CEO Francis Irving says the scraper is easy to use. “Once you know the search term or user you want to archive you can create the dataset and let the tool run. Once you have enough tweets, you can download the data for analysis,” he said via Skype. The software can also visualize and summarize the data, saving much work for the researcher.

A similar freemium tool is Simply Measured's **RowFeeder**, but the datasets are more expensive and the results are more limited than ScraperWiki’s. For example, RowFeeder’s cheapest product, which costs $35 a month, includes just three datasets for a maximum of 5,000 tweets per month.

In the absence of an archive from the Library of Congress, ScraperWiki seems a reasonable solution to the ongoing problem of how to collect and analyze meaningful Twitter data. Its relatively inexpensive pricing partially addresses the growing two tiers in academic research caused by the high cost of data analysis. As Hermida said in the survey, "Paying for access means that there could be a divide between data-rich and data-poor researchers."

*Correction: A previous version of this story misspelled Mike Reilley's name.*
Kelly Fincham, an assistant professor at Hofstra University, has been using Twitter for research since 2010.

Tags: Twitter