Collecting and archiving tweets: a DataPool case study

Steve Hitchcock
JISC DataPool Project, Faculty of Physical and Applied Sciences, Electronics and Computer Science, Web and Internet Science, University of Southampton, UK
Version 2.0, 16 May 2013

Abstract

Information presented to a user via Twitter is variously called a 'stream', that is, a constant flow of data passing the viewer or reader. Where the totality of information passing through Twitter at any moment is considered, the flow is often referred to as a 'firehose', in other words, a gushing torrent of information. Blink and you've missed it. But does this information have only momentary value or relevance? Is there additional value in collecting, storing and preserving these data?

This short report describes a small case study in archiving collected tweets by, and about, a research data project, DataPool at the University of Southampton. It explains the constraints imposed by Twitter on the use of such collections, describes how a service for collections evolved within these constraints, and illustrates the practical issues and choices that resulted in an archived collection.

The second version of the report adds a short postscript on rights, ethics and privacy of archiving Twitter data, prompted by a Twitter dialogue on this report. Two additional references of related work at Southampton are provided towards the end of sections 1 and 2.

1 Preserving, delivering and reusing tweets: Twitter rules

At one level, that of preserving the dialogue and evolving culture of a massively popular service, yes there ought to be value in archiving selected data from Twitter. The Library of Congress (LoC) in the USA already collects and preserves the entire Twitter output. Here are some numbers to gauge the scale of that operation: LoC had by January 2013 received 170 billion tweets totaling 133.2 terabytes for two compressed copies. Each day the Library receives nearly half a billion tweets, as of October 2012, up from 140 million/day in February 2011 http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

So you think a tweet is just 140 characters? In the LoC view each tweet has more than 50 accompanying metadata fields, such as place and description, date and time, number of followers, account creation date, geodata, etc. Each tweet is a JSON file. This is not specific to LoC. We will learn more about this format later.

Formative studies by Social Media in Live Events (SMiLE) of social media activity, including Twitter as well as video and photo services, surrounding an event have shown the value of collecting such data as well as the need for real time collection. In addition SMiLE has begun to reveal the ethical implications of archiving and mining social media content.

If you don't have social media, you are no one: How social media enriches conferences for some but risks isolating others http://blogs.lse.ac.uk/impactofsocialsciences/2012/05/23/social-media-enrich-but-isolate/

Beyond the special relationship with LoC, Twitter, a commercial entity seeking to transform popularity and usage into commensurate revenue and profit, has not always assisted the process of collection, preservation and reuse, with the particular sticking point being reuse. For organisations dedicated to preservation, this is not new.

As Twitter has sought to establish control over content delivered via its service, third-party application developers have been constrained by terms and conditions applying to content interfaces and reuse. In particular, Twitter discourages, though does not prohibit, developers from creating applications that "mimic or reproduce the mainstream Twitter consumer client experience." http://arstechnica.com/business/2012/08/new-api-severely-restricts-third-party-twitter-applications/

One Twitter service and preservation provider that fell foul of more vigilant enforcement of Twitter's terms was TwapperKeeper (TK), used by many academics and researchers. This service allowed users to set search terms for TK to collect and store tweets. These collections could be searched and accessed by other users on the Web.

Twitter rules restrict not the collection of this data but its redistribution by a third-party service such as TK, at least, redistribution in a form that looked like a Twitterstream. This was the relevant paragraph in the Twitter T&Cs:

I.4a. "You will not attempt or encourage others to: sell, rent, lease, sublicense, redistribute, or syndicate the Twitter API or Twitter Content to any third party for such party to develop additional products or services without prior written approval from Twitter."
Twitter Developer Rules of the Road https://dev.twitter.com/terms/api-terms

There is more, and we will return to this point in the section on archiving tweets. To begin to understand how this limitation might affect archiving, we need to differentiate the Twitter data architecture and interaction model. In Tinati, et al's illustration the firehose of tweets is just one element of the data architecture.

Tinati, Ramine, Carr, Leslie, Hall, Wendy and Bentwood, Jonny (2012) Identifying Communicator Roles in Twitter. ePrints Soton, 10 Mar 2012. In *Mining Social Network Dynamics,* Lyon, Apr 2012
http://eprints.soton.ac.uk/335268/

TwapperKeeper fell foul of the rule. While conforming and closing the infringement early in 2011, it became clear that since launching in 2009 TK had induced users to enjoy a type of service it could no longer support. In January 2012 TK became part of the Hootsuite social media management services.

Twitter Cracks Down on Another Third-Party App, TwapperKeeper
http://www.pcmag.com/article2/0,2817,2380784,00.asp

2 Tweepository: an EPrints app and repository

It was in this context that in 2011 work by Adam Field with EPrints repository software at the University of Southampton led to an EPrints application called Tweepository. Like TK, this collects and presents tweets based on a user's specified search term, for a specified period of time. Collections are stored as items within an EPrints repository. The tweets can be viewed as if they were a tweetstream, with various data analysis tools - Top Hashtags, Top Tweeters, Top Tweetees, Top Links, and a weekly count of tweets - and results are displayed in the same view.

Using EPrints Repositories to Collect Twitter Data
http://repositoryman.blogspot.co.uk/2011/10/using-eprints-repositories-to-collect.html

Tweepository can be found in the EPrints Bazaar (http://bazaar.eprints.org/267/) or app store , and applied to an installed version of EPrints (version 3.3 or higher can be used with apps) by a repository administrator. At Southampton, after running a number of test repositories with the app installed, in late 2012 Field set up a new, institutionally supported beta instance of a repository for Twitter collections rather than add this functionality to an existing repository at the university.

Collections can be initiated by creating a record in a modified EPrints deposit form (Figure 1). Collections must have a search parameter and set an expiry date. The collection begins from the date the record is created but will retrospectively pull in items from the search from up to a month prior to the date of creation, depending on the number of retrospective items, again acting within constraints set by Twitter.

Adhering to current Twitter terms of use for such services, only registered users can login and view these collections. The collections can be exported in a range of archival formats using designated tools, but 'programmatic' formats, from which the original look-and-feel and services might be recreated, cannot be openly redistributed. In other words, only human read-only versions may be viewed by archivists and other retrospective users of the collected and archived content.

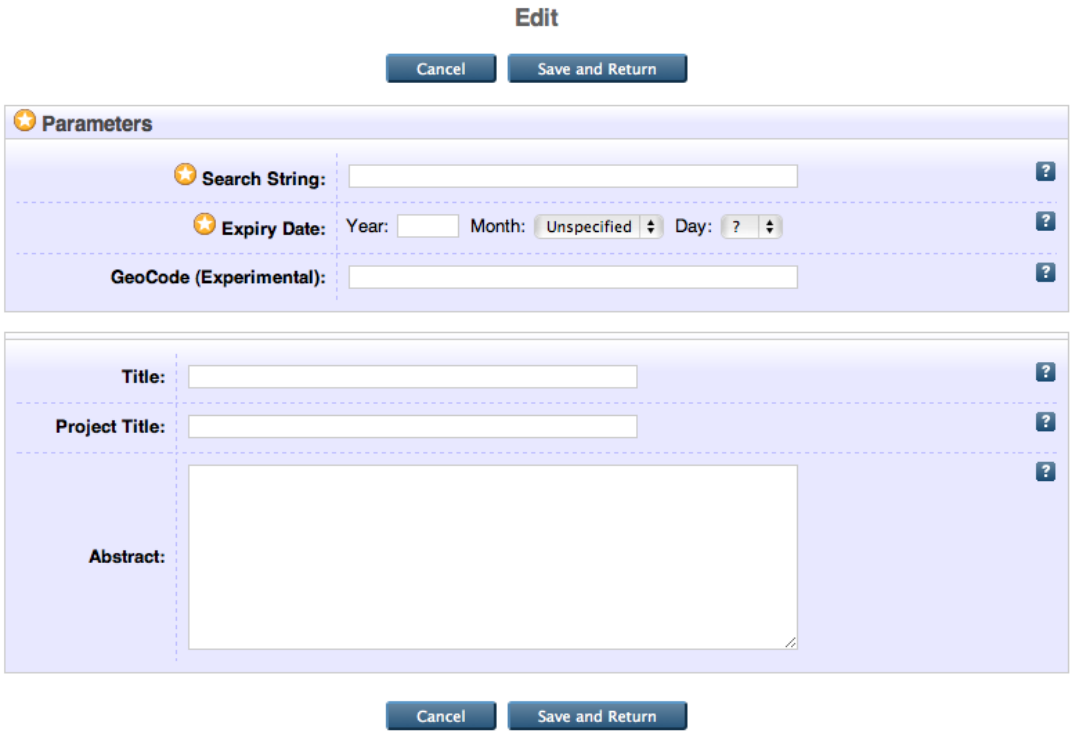

Figure 1: Creating and editing a record for a Tweepository collection.

AryaArjmand provides more illustration of the harvester in use at Southampton.

Amir AryaArjmand, Twitter harvesting, Digital Economy USRG, December 14, 2012 http://digitaleconomy.soton.ac.uk/blog/2975

3 DataPool: archiving tweets as research data

As an institutional research data project, DataPool (http://datapool.soton.ac.uk/) is investigating how the University of Southampton can provide services for managing the data outputs of researchers. Like many similar institutional projects within the JISC Managing Research Data Programme 2011-13 (http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx), DataPool opened a Twitter account (on 22 November 2011) to record progress and communicate this within the community. In this way DataPool is a data producer too, and has a responsibility to set an example for managing and storing these data. It was written into the project plan that we would explore ways of archiving outputs such as our Twitter stream.

The project and researchers associated with the project were among early users of Tweepository using the app installed on a test repository. Collections began in August 2012 based on the project's Twitter user name, other user names, and selected hashtags, from conferences or other events (Figure 2).

Figure 2: One collection among many, Tweepository file record for jiscdatapool collection, allowing the registered user to view or edit the file, or package for export.

As the collections grew in number and size the institutional Tweepository mentioned above was launched in late 2012. A packager tool provided by Adam Field enabled the ongoing test collections to be transferred to the supported Southampton Tweepository without a known break in service or collection. The packager (the brown icon seen in Figures 2 and 3) creates a zip file containing sets of CSV and JSON files.

Figure 3: Example collection, tweets by or on JISC DataPool Project using search term jiscdatapool, presented by University of Southampton beta Tweepository. In the left-hand column are items from the tweetstream (300+ tweets from 10 August 2012, set to conclude on 31 March 2013); in the right-hand column are results of data analysis of tweets.

Our jiscdatapool collection in Tweepository is from August 2012, the start of the test collection; DataPool began using Twitter in November 2011. Tweepository, and other Twitter harvesting services, are limited in the extent of retrospective searches permitted by Twitter, as noted above, so we are unable to retrieve earlier tweets using this service. If completeness is important, collections must begin promptly with the first occurrence of the search term.

As of December 2012 Twitter makes it possible for users to download their Twitter archive, containing all tweets (including retweets) from the beginning of the account (http://blog.twitter.com/2012/12/your-twitter-archive.html). This archive is in the form of a zip file containing CSV and JSON, i.e. just as the Tweepository Packager tool. Tweets are stored in files per year and month. This service only applies to the archive of a registered Twitter user, not the general search collections possible with Tweepository.

4 Archiving Tweepository collections

The Southampton Tweepository remains in a beta form, implying no guarantee is provided about the continuity of the service, nor of a permanent archive. It is, however, supported by the IT services provider across the university, iSolutions, and could be seen as a precursor to institutional commitment to the service on completion of a successful test and monitoring phase.

For completeness as an exemplar data case study, given that institutional services such as Tweepository are as yet unavailable elsewhere, towards the end of the DataPool project in March 2013, tweet collections were archived. We used the provided export functions to create a packaged version of selected, completed collections for transfer to another repository at the university, ePrints Soton http://eprints.soton.ac.uk/, which has a commitment to long-term storage of deposited items.

Figure 3 shows the formats available to export the collection to another service. As well as the Twitter packager, for the registered administrator of the collection the full list is:

- CSV
- EP3 XML
- HTML
- JSON
- JSON TweetStream
- Wordle Link

For archival purposes we choose not to use system exporters such as EP3 XML and JSON; nor Wordle, as this is a specific derived view on the data.

Most suitable formats, according to Field, appear to be:

- CSV: “All the tweets, with the most important bits of data (though some

bits are missing)"

- JSON TweetStream: "All the data in a form optimised for programming, but vaguely human-readable"

JSON is a data interchange format favoured by Twitter, and as we have seen, is used for the LoC collection.

Here we need to return to Twitter T&Cs, point I.4A, because there is more:

> "You may export or extract non-programmatic, GUI-driven Twitter Content as a PDF or spreadsheet"
> https://dev.twitter.com/terms/api-terms

Our interpretation of this is that archived collections based on CSV and JSON could be archived but not made openly available. Only PDF is derivable from the original collection and falls within the term above for being made openly available. PDF can be created from an HTML Web page using a standard 'Save as PDF' function on any computer.

In this case we have the original Tweepository view of the collection (Figure 3), with the data analysis tools (right-hand column) shown but a significant section of "tweets not shown..." between the earliest and latest tweets. Or a Web page created with the HTML export option provided in Tweepository produces a chronological list of all the saved tweets, an extended version of the view shown in Figure 3, with the "tweets not shown..." returned, but omits the results provided by the data analysis tools.

Thus attached to our archived tweet collections in ePrints Soton (Figure 4) are:

1. Reviewable PDF of the original Tweepository Web view (Figure 3) with some "tweets not shown..."
2. Reviewable PDF of complete tweet collection without data analysis, from HTML export format

3. JSON Tweetstream* saved using the provided export tool
4. Zip file* from the Packager tool

* reviewable only by the creator of the record or a repository administrator

For completeness we have added the zip archive downloaded directly from Twitter, spanning the period from opening the account in November 2011.

Figure 4: file-level management in ePrints Soton, showing the series of files archived from Tweepository.

5 Conclusion

Tweepository harvests, stores, displays and exports tweets collected according to specified search terms and time periods, within the scope of current, limiting terms and conditions imposed by Twitter. Twitter-like functionality can be provided for the collections, but only to the creators of the collection and other registered users of the Tweepository instance. Where those collections are exported and reused this is in non-Twitter formats providing human-readable but non-programmable views of the collections.

DataPool has created and exported archival versions of its tweet collections in permitted formats from Tweepository to the University of Southampton EPrints repository. This provides some assurance of longer-term storage of deposited

items than the current beta Tweepository, but cannot provide the same functionality or views.

What value the data in these collections will prove to have will be measured through reuse by other researchers, and remains an open question, as it does for most research data entering the nascent services at institutions such as the University of Southampton.

At LoC researchers do not yet have access to its Twitter archive, and this is now a 'priority' for the Library. "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."
http://www.wired.co.uk/news/archive/2013-01/07/library-of-congress-twitter

Archiving tweets is a first step; realising the value of the data is a whole new challenge.

6 Postscript: rights, ethics and privacy of archiving Twitter data

A blog post summary of the first version of this report prompted a Twitter-based dialogue with Robin Rice and Brian Kelly on the issues of rights, ethics and privacy of archiving Twitter data, later added in two comments to the blog post http://datapool.soton.ac.uk/2013/03/27/collecting-and-archiving-tweets-a-datapool-case-study/

It is true this report does not consider these issues in any detail, and the case was made for a more detailed consideration, especially with respect to privacy. Robin Rice highlighted a solid review of these issues provided by Small, et al.

Heather Small, Kristine Kasianovitz, Ronald Blanford, Ina Celaya, What Your Tweets Tell Us About You: Identity, Ownership and Privacy of Twitter Data, *International Journal of Digital Curation*, Vol. 7, No. 1, 2012
http://ijdc.net/index.php/ijdc/article/view/214

Although Small et al. make reference to the framework provided by Twitter for collecting and archiving tweets, in terms of its two APIs, and the Twitter terms of service for reuse of these data, they do not examine the limitations these conditions impose on the presentation of an archive - which this report addressed - and the natural constraining effects these have on the issues under consideration, which are significant.

While it is easy to claim that most tweets are public by design and intent, this does not in itself justify collecting and providing access to an archived collection. Tweets are sent with the expectation of instant publication within a fast-moving stream of information, but without indication from the tweeter that the content is intended to have any longer-term use. "One feature of social media is that communications can be short in length and short in lifespan" (Darling, et al.) Archiving challenges that lifespan assumption (although not necessarily in the case examined by Darling, et al.).

Emily S. Darling, David Shiffman, Isabelle M. Côté, Joshua A. Drew, The role of twitter in the life cycle of a scientific publication, arXiv, 2 May 2013 http://arxiv.org/abs/1305.0435

Small, et al., report resistance, in privacy terms, to the LoC Twitter archive, notably in comments to LoC's FAQ about the service and, apparently, a pre-archiving tweet deletion service NoLoc.org, now expired.

The use case considered in this report of archiving a project Twitter archive is tiny in comparison with anticipated Twitter collections such as at the LoC, yet even at this scale and level of quantifiability it raises concern. Cases need to be considered not just in terms of the archived collection and collection policy, as recommended by Small, et al. (based on Charlesworth), but also in terms of presentation and access, target users and purposes, and the tools available for users to interrogate the archive.