Provenance OF A TWEET

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

NATIONAL ARCHIVES
NATIONAL HISTORICAL PUBLICATIONS & RECORDS COMMISSION

Authors: Dan Kerchner, Justin Littman, Christie Peterson, Vakil Smallen, Rachel Trent, Laura Wrubel

Read the full text http://bit.ly/prov-tweet-doc
http://go.gwu.edu/sfm ● @SocialFeedMgr ● sfm@gwu.edu

Abstract

Social Feed Manager (SFM) is an open-source tool for collecting social media data via the APIs for services including Twitter, Flickr, Weibo and, soon, Tumblr.

The data retrieved via the API is significantly richer than what is typically seen by website or app users.

This data, combined with data captured or generated by SFM and data entered by collectors allows us to gather provenance information about three activities in the life of a Tweet:

- creation
- collection
- selection

Questions remain about which of these categories are most useful, and which should be privileged for capture and display.

Background

SFM was developed by a team at George Washington University (GW) Libraries and has successfully supported GW scholars collecting tweets for their research since 2012.

Social media researchers have called for increased documentation to support reproducibility, validity and data sharing now and in the future. These needs align closely with the concept of provenance as used in archives and digital curation.

Provenance **is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.**

PROV Overview, W3C Working Group Note 30 April 2013

Creation

The Twitter API includes metadata about the Twitter user who authored the tweet:

```
"user": {
    "id": 216776631,
    "verified": true,
    "description": "Join our campaign for president at
https://t.co/nuBuf1GIwb. Tweets by staff.",
    "location": "Vermont",
    "screen_name": "BernieSanders",
    "lang": "en",
    "name": "Bernie Sanders",
    "url": "https://t.co/W6f7Iy1Nho",
    "created_at": "Wed Nov 17 17:53:52 +0000 2010",
    "time_zone": "Eastern Time (US & Canada)",
},
```

how it was posted:

```
 "source": "<a href=\"https://about.twitter.com/products/tweetdeck\"
rel=\"nofollow\">TweetDeck</a>",
```

when it was posted:

```
"created_at": "Fri May 20 16:02:03 +0000 2016",
```

and possibly the country name and geolocation where it was posted.

Collection

For each call to the API, SFM records the *exact HTTP transaction* as WARC request and response records.

WARC request records include:

- URL from the API that was retrieved
- query parameters that were passed

```
GET
/1.1/statuses/user_timeline.json?count=200&max_id=682336123457155073&
user_id=216776631 HTTP/1.1
```

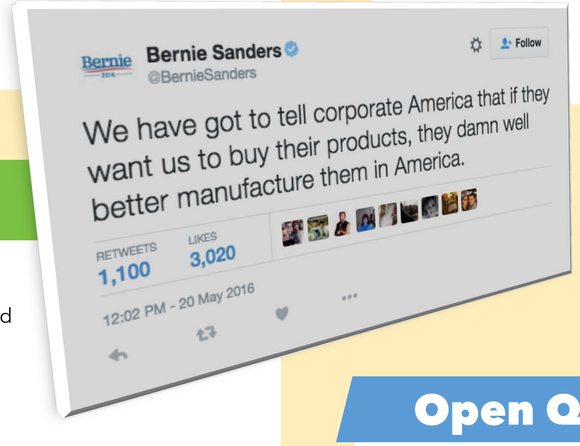
WARC response records include the content of the tweets as JSON as well as data about the API response, such as

- server status code
- date stamp for when the data was collected
- information on rate limits that could affect what is collected

The headers on the WARC records hold metadata about how each message was sent or received, including a date stamp, fixities for the HTTP message, and server IP address.

Additionally, SFM captures its own metadata on collection activity, which is recorded in a database record. This includes

- dates collection started and ended
- some basic statistics
- any informational, warning, or error messages
- token or uid updates received



Selection

Change log

Date	User	Fields
May 24, 2016, 1:51 p.m.	justinlittman	schedule_minutes: "10080" changed to "1440" **Note:** Given the volume of tweeting, increasing collection frequency.

SFM automatically logs all actions taken on a collection.

Additionally, SFM gives collectors the ability to add notes to the change log to document curatorial activities at key moments in the collecting process:

- creating a collection
- organizing a collection
- adding or removing seeds
- changing the schedule, frequency or other parameters for collecting
- changing the credentials used for collecting

Open Questions

- Which of this provenance metadata is most valuable to researchers, archivists or other collectors?
- What prompts should we provide to encourage collectors to document the provenance of their collections?
- How should we provide access to provenance metadata?
  - In SFM's UI?
  - In reports when exports are created?
  - Exposed via SFM's software libraries?
  - A REST API?
  - Machine-readable? Human-readable?
- Is there value in a mapping to or providing output in accordance with metadata standards such as PREMIS or PROV?
- Are there other relevant specifications or standards that we should consider?

SFM development was supported from 2013-2014 by a grant from the Institute of Museum and Library Services

This work is supported by grant #NARDI-14-50017-14 from the National Historical Publications and Records Commission