Archiving Social Media
A Case Study

Megan O’Hern @megan_ohern
Amy Wickner @amelish

8th Annual Bridging the Spectrum Symposium • Catholic University of America • February 12, 2016
#cualis16 #S3pm
Background
Born-Digital Strategy

Collection contexts:

- Transferring files
- Imaging digital media
- Web archiving
- Preserving email accounts
- Harvesting social media

Shared concerns:

- Privacy
- Ownership
- Significant properties
- Access conditions
- Storage
Why Twitter?

- AFL-CIO uses Twitter as “soft official communication”
- Opportunity for macro-appraisal
- Opportunity for documentation strategy
- Opportunity for democratization of archives and the elevation of participants

#cualis16 #S3pm • @megan_ohern • @amelish
Uses of Twitter Data

- Many different types of analysis:
  - Content analysis
  - Network analysis

- Breadth in applicability:
  - Social justice
  - Business
  - Emergency management
  - Citizen science
Preserving Dynamic Digital Material

- “Significant properties” of social media
- Formal v. functional preservation
- Preserving versions and variability
- Media independence: is it possible?
Challenges

- What does appraisal look like for Twitter?
- How to integrate with born-digital policies?
- Collection development priorities
- Navigating Twitter’s API
- How much technical knowledge is needed?
Tools
Twitter’s Personal Archive

Most Useful For:

- Archiving a complete archive feed
- Archiving an account you own

Least useful for:

- Capturing interaction and behavior
- Gathering lots of metadata about a tweet
Most useful for:

- Crawling single or few Twitter feeds as part of a web archiving program

Least useful for:

- Capturing more than 24 hours of tweets
- Capturing tweet metadata

#cualis16 #S3pm • @megan_ohern • @amelish
Twitter Archiving Google Sheet (TAGS)

Most useful for:

- Creating a snapshot of Twitter activity around a search term
- Creating simple network visualizations

Less useful for:

- Creating a complete archive of an account/search term
- Automatically collecting data
- Maintaining privacy
With this spreadsheet you can:
- automatically pull results from a Twitter Search into a Google Spreadsheet

Instructions:
1. If not already File > Make a copy this template while logged into your Google account
2. If there is no TAGS menu click this button --> Enable custom menu
3. If you've never run TAGS > Setup Twitter Access do so now (this should only need to be done once for all your TAGS sheets)
4. Enter term <aficio> -- you can use search operators like AND OR as well as from: and to: eg: #JobsNow AND from:BarackObama (without quotes)
5. Make a one off collection with TAGS > Run now or set a trigger to collect every hour TAGS > Update archive every hour. To change the frequency open Tools -> Script Editor then Triggers -> Current scripts' triggers... and adjust

Advanced Settings:
- Period: default
- Follower count filter: 0
- Number of tweets: 3000
- Type: search/tweets

Stats
- Number of Tweets: 0
- Unique tweets: 0
- First Tweet: 29/01/2016 23:27:24
- Last Tweet: 04/02/2016 20:54:42
<table>
<thead>
<tr>
<th>from_user</th>
<th>text</th>
<th>created_at</th>
<th>time</th>
<th>geo_coordina</th>
<th>user_lang</th>
</tr>
</thead>
<tbody>
<tr>
<td>VioletL_Phoenix</td>
<td>RT @FlushtheTPP: TODAY! Cities rise to #StopTPP #AntiTPP #TPPNoWay see <a href="https://t.co/dhXntTNRw1">https://t.co/dhXntTNRw1</a> @AFLCIO @Teamsters @TPPMediaMarch https://...</td>
<td>Thu Feb 04 02/04/2016 04:54:42</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>delane392</td>
<td>@CarmenSpinDiego to speak at the South Carolina #RaisingWages Summit 2/6 Learn more <a href="https://t.co/641FrXiA2x">https://t.co/641FrXiA2x</a> #1u <a href="https://t.co/HTgtmLgZks">https://t.co/HTgtmLgZks</a></td>
<td>Thu Feb 04 02/04/2016 05:53:35</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>delane392</td>
<td>@Tefere_Gebre to speak at the South Carolina #RaisingWages Summit on 2/6. Learn more: <a href="https://t.co/641FrXiA2x">https://t.co/641FrXiA2x</a> #1u <a href="https://t.co/y5E6YIS4j5">https://t.co/y5E6YIS4j5</a></td>
<td>Thu Feb 04 02/04/2016 05:52:16</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>delane392</td>
<td>Proud to welcome @ChrisKromatic to the SC Summit on #RaisingWages. Join us. Register: <a href="https://t.co/641FrXiA2x">https://t.co/641FrXiA2x</a> #1u <a href="https://t.co/0hGHDd1xCR">https://t.co/0hGHDd1xCR</a></td>
<td>Thu Feb 04 02/04/2016 05:51:27</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>delane392</td>
<td>Proud to welcome @ClyburnSC06 to the SC Summit on #RaisingWages. Join us. Register: <a href="https://t.co/641FrXiA2x">https://t.co/641FrXiA2x</a> #1u <a href="https://t.co/iFYNuUuRnL">https://t.co/iFYNuUuRnL</a></td>
<td>Thu Feb 04 02/04/2016 05:50:12</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>delane392</td>
<td>Proud to welcome @GCobbHunter to the SC Summit on #RaisingWages. Join us. Register: <a href="https://t.co/641FrXiA2x">https://t.co/641FrXiA2x</a> #1u <a href="https://t.co/LonlyIHx9Z">https://t.co/LonlyIHx9Z</a></td>
<td>Thu Feb 04 02/04/2016 05:49:39</td>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChezypeepQueen</td>
<td>RT @Teamsters: 5 Reasons that Rosy Predictions About the #TPP Are Wrong: <a href="https://t.co/leW9jflmC1">https://t.co/leW9jflmC1</a> #StopTPP <a href="https://t.co/AzURNjYuhl">https://t.co/AzURNjYuhl</a></td>
<td>Thu Feb 04 02/04/2016 05:47:28</td>
<td>en</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NodeXL

Most useful for:

- Creating network adjacency lists and visualizations
- Archiving interactions on Twitter

Least useful for:

- Creating a complete archive of a user/search term
- Automatically collecting data
WebRecorder

Most useful for:

- Capturing dynamic web content
- Documenting user experiences
- Archiving by browsing or scrolling
- Capturing WArc files for ingest into a digital repository

Least useful for:

- Capturing online content with large number of pages
Most useful for:

- Collecting tweet data via Twitter search or filter stream APIs, hydrating via lookup API
- Transforming data for visualization

Least useful for:

- Capturing the look and feel of Twitter
Social Feed Manager

Most useful for:

- Managing the collection of Twitter data related to multiple accounts or hashtags over time
- Offering tweets as research data

Least useful for:

- Capturing the look and feel of Twitter

@cualis16 #S3pm • @megan_ohern • @amelish
@LizShuler: Labor has a lot of work to do around issues of race & economic justice. This is a start https://t.co/oRzShpdr8b #1uMLK #1uBHM
Access
Collection Overview

**Title:** 2014 US-Cuba Policy Change Twitter Archive, 2014-2015

**ID:** CHC5404

**Creator:** University of Miami

**Extent:** 21.0 Gigabytes

**Arrangement:** The data is available as one downloadable zip file.

**Languages:** English, Spanish, Castilian

Abstract

The 2014 US-Cuba Policy Change Twitter archive collection contains a data set of tweets collected from the Twitter microblogging platform when President Barack Obama announced on December 17, 2014 that the United States would begin normalizing full diplomatic relations between the U.S. and Cuba after more than half a century of minimal relations. The Cuban Heritage Collection collected tweets relating to the hashtags #cuba, #cubapolicy, #cubalibre, #cubusa, #uscuba, and #cubanmiami.

Scope and Contents of the Materials

The 2014 US-Cuba Policy Change Twitter archive collection contains a data set of tweets collected from the Twitter microblogging platform when President Barack Obama announced on December 17, 2014 that the United States would begin normalizing full diplomatic relations between the U.S. and Cuba after more than half a century of minimal relations. President Obama's announcement included plans to re-establish the US embassy in Havana, allowing official visits of Cuban diplomats and officials to the United States, and increased official dialogue on public policy issues affecting both countries.

During the announcement, the Cuban Heritage Collection collected tweets relating to the hashtags #cuba, #cubapolicy, #cubalibre, #cubusa, #uscuba, and #cubanmiami between December 9, 2014 and January 28, 2015. The tweets collected before the December 17, 2014 announcement by Presidents Obama and Castro represent tweets gathered via the Twitter API seven days before the announcement.

Administrative Information

**Access Restrictions:** Due to Twitter’s Terms of Service, this data archive is accessible only to the University of Miami community. It can be accessed via an IP-restricted portal in the University of Miami Scholarly Repository. For specific questions, please contact chc@miami.edu.

**Use Restrictions:** Due to Twitter’s Terms of Service, this data archive is only accessible onsite at the University of Miami.

**Acquisition Method:** This collection was acquired and collected by the University of Miami Libraries using the Python program Twarc, developed by Ed Summers.

**Related Materials:** Selected websites have been archived by the University of Miami’s Cuban Heritage Collection relating to the US-Cuba policy change announcements. They are accessible online: [https://www.archive-it.org/collections/5163](https://www.archive-it.org/collections/5163)

**Preferred Citation:** 2014 US-Cuba Policy Change Twitter archive, Cuban Heritage Collection, University of Miami Libraries, Coral Gables, Florida.

**Processing Information:** This collection was a collaboration between several University of Miami Libraries departments, including the Cuban Heritage Collection, Web and Emerging Technologies, and Digital Strategies. Project members include: Natalie Baur, Jaime Little, Andrew Darby, Sevika Singh, Sarah Shreeves, Lyn MacCorckle, Laura Capell, and Jason Cohen.
Guide to the Stanford University Twitter Collection SC1205

Collection Title: Guide to the Stanford University Twitter Collection SC1205
Collection Number: SC1205

Description: The Stanford University Twitter Collection, 2009-2014, includes csv files and html/CSS files for the @Stanford and @StanfordNow Twitter feeds, as well as a sample weekly report on Stanford "mentions" on Twitter over a one-week period in 2014, generated through Salesforce/Radian6.
Extent: 13.0 megabyte(s) (228 files in 56 folders)
Restrictions: All requests to reproduce, publish, quote from, or otherwise use collection materials must be submitted in writing to the Head of Special Collections and University Archives, Stanford University Libraries, Stanford, California 94305-6064. Consent is given on behalf of Special Collections as the owner of the physical items and is not intended to include or imply permission from the copyright owner. Such permission must be obtained from the copyright owner; heir(s) or assigns. See: http://library.stanford.edu/spc/using-collections/permission-publish.
Availability: The materials are open for research use. Audio-visual materials are not available in original format, and must be reformatted to a digital use copy.
Title: Smithsonian Feed (SmithsonianFeed) on Twitter

URL: http://twitter.com/SmithsonianFeed/

Description: Smithsonian Institution Social Media Smithsonian Institution Social Media

Captured 4 times between Dec 3, 2012 and Jan 9, 2013

Subject: Website, Web sites

Creator: Smithsonian Institution

Publisher: Smithsonian Institution

Language: ENG, ENG

Date: 2012-12-03

Collector: Smithsonian Institution Archives, Smithsonian Institution Archives

Rights: Restricted, Restricted
This item represents a collection of 13,480,000 tweet IDs that mentioned "ferguson" from 2014-08-10 to 2014-08-27 and 15,080,078 tweet IDs that mention "ferguson" between 2014-11-11 and 2014-12-08.

The first set includes tweets for the three week period after the shooting of Michael Brown, and the second range includes tweets around the grand jury’s decision not to indict police officer Darren Wilson which was announced on 2014-11-24.

The first set of tweets was collected by Ed Summers at the University of Maryland and the second was a collaboration between Molly Loyd, Gregory Coleman, Kimberly Lamke, Benjamin Sugar and Ed Summers.

At the time of writing (January 7th, 2015), Twitter’s Terms of Service do not allow bulk distribution of the raw JSON data for tweets. However they do allow the tweet IDs to be shared. You can use this data in combination with their data API to “hydrate” the IDs to get the original data. A tool like tweet is can be used to hydrate them automatically.
What’s Next?
Thoughts Going Forward

- Preserving social media?
- Lowering barriers to entry
- Changing nature of social media
Tools and Resources

- Twitter Personal Archive
- ArchiveIt
- twarc
- TAGS
- NodeXL
- WebRecorder
- Social Feed Manager

- NCSU Social Media Archives Toolkit
Research with Twitter

- Ferguson Twitter Archive
- “Tweet Earthquake Dispatch: Crowdsourcing Earthquake Detection”
- Quinn, B. and J. Ball (8 July 2014) “US military studied how to influence Twitter users in Darpa-funded research.” *The Guardian*
Archivists’ Perspectives


