PARC:

LIBRARY OF DIGITAL POLITICS
DIGITAL ARCHIVE OF POLITICAL PROTESTS

FACULTY OF INFORMATION

ALEXANDER HERD
MAGDELENE SCHIFFERER
MY AHN TRUONG
KALI BRANDEN
BRIAN LAU

PARC (Political ARChive)

is our envisioned library of digital politics. It serves as a digital archive of political protests, in particular for political tweets, and is staffed by the *Political Tweets Librarians*. PARC is an extension of the *guidelines* we developed for the archiving of political Tweets and is an attempt at addressing the issues of archiving increasingly ephemeral digital knowledge and the consequent changing role of libraries in society.

ARC
CO.PE
CURATE

Digital political knowledge is in danger of disappearing. This is especially pertinent to the poltical Tweet. The political Tweet is an important facet of our political culture and history so it is imperative that archivists and librarians archive, harvest and curate Tweets related to international political events for the purposes of *future* research, study and education.

PARC is calling on archivists and librarians to protect this fragile facet of our political knowledge. We here offer a set of guidelines on how to begin.

The U of T Robarts Library is a participant in an international effort to archive, harvest and curate political Tweets and other digital political knowledge. If you are interested in playing a part, please contact :

Sam-chin Li, Reference and Government Publications Librarian **samchin_li@utoronto.ca**

ARC
Using Archive-It for Tweets

- Ensure that Twitter feed is publicly available (default setting for most Twitter accounts).
- Select Twitter feed as seed URL, adding an ending '/' to the URL, to ensurecapturing only the specific feed, not all of Twitter.
  - For example, http://twitter.com/internetarchive/ (with an ending /)
  - Seed URLs cannot have a 'www' – by default, Twitter URLs do not have a www and www.twitter.com is blocked by 'robots.txt'.

§ Reminder: To enable the 'Ignore Robots.txt' feature, use the Archive-It 'Submit a Question' link to have the Archive-It Support team turn on this feature for the account.

- By default, dynamically scrolling content should be in scope and captured.
- Languages: For any given Tweet, the page is captured in all languages the Twitter interface supports, including Chinese. For example, for an original Tweet URL of the following format:
  http://twitter.com/[user name]/[tweet ID]/

The following are also captured:
http://twitter.com/[user name]/[tweet ID]?lang=zh-tw (Traditional Chinese)
http://twitter.com/[user name]/[tweet ID]?lang=zh-cn (Simplified Chinese)

- To capture Twitter content in only one language, use a 'regular expression' to block other language variations.

On the Archive-It account:

- From the 'Host Constraints' tab of the 'Modify Crawl Scope' page, add the host 'Twitter.com'.
- Then click 'Add Rule'; from the dropdown menu select 'Matches the following regular expression'.
- In the box add the following: ^.*lang=(?!en).*$ (to capture English only).
- Adjust the regular expression to capture in other languages by changing the language abbreviation in the parenthesis; for example, to capture only Traditional Chinese content, use ^.*lang=(?!zh-tw).*$
- **Run a test crawl after adding this regular expression.**

Other notes when using Archive-It:

- Robots.txt might block styling and image information: in these cases, ignore robots.txt for twimg.com (Twitter's photograph hosting domain).
- Crawling links in Tweets: Many Tweets contain a URL that links to a page outside of Twitter (often using a URL shortener service such as bit.ly or tinyurl.com); by default, these links from Twitter posts will not be captured, but this can be done easily, as all links in Tweets redirect through the Twitter URL shortener: http://t.co/
  - To include the pages linked on Tweets in a crawl, add an 'Expand Scope' rule for any URLs that contain http://t.co/ to allow the crawler to go to the link (or in the case of URL shorteners, the URL that the shortener redirects to), and archive that page (i.e., the actual page and content that was Tweeted about).

§ All embedded files (images, CSS files, javascript files, etc.) will also be archived, enabling a complete capture of the page Tweeted about.

§ More than this one page will not be crawled (unless there are additional scoping rules in place), so an entire site that is linked to in Tweet will not be crawled.

§ Again, conduct a test crawl with this new Expand Scope rule in place, to ensure it has been entered properly.

- t.co URLs are blocked by robots.txt; therefore, ignore robots.txt for t.co.
- This Expand Scope rule will apply to all seeds and crawls in the collection.

- Archiving Twitter searches: Search pages should archive similarly to Twitter feeds, however dynamically scrolling content on search pages may not be captured.
  - As with Twitter feeds, include an ending '/' on any search seed URLs to avoid putting all of Twitter in scope.

§ For example: twitter.com/search?q=Michi gan%20right-to-work&src=typd/

(Sources: Bragg & Rollason-Cass, 2014; Twitter, Inc., 2014b)

CO.PE
Copyright, Privacy and Ethics

Copyright

- Copyright and social media is an emerging issue and as of yet has not been tested by law or written into legislation.
- Scholars argue either for and or against Tweets' eligibility for copyright.
- Twitter's Terms of Service state that by posting Tweets, authors consent to their Tweets' collection and archiving.
- Political Tweets are exempt from copyright under the Fair Dealing exemption, Copyright Act of Canada, because they are used for education, study, or research.
- Develop an official document stating who, when, where, and what access is granted as part of institutional social media curation policy.
- A dark archive is recommended to place Tweets for a minimum of 75 years to protect the institution from copyright issues that accompany regular public access.

(Sources: Antracoli, et al., 2014; Copyright Act, 1985; Small, et al., 2012; Twitter, Inc., 2014a)

Privacy and Ethics

One of the biggest questions is how to address ethical concerns surrounding the harvesting and curation of Tweets. This question is addressed by four broad privacy and ethical guidelines, based on the core principles in the American Anthropological Association's Statement of Ethics (2012):

1) Protect Twitter Authors From Any (Direct and Indirect) Potential Harm.

2) Ensure to the Best of Your Knowledge the Author has Granted Consent to Capture and Curate Twitter Data.

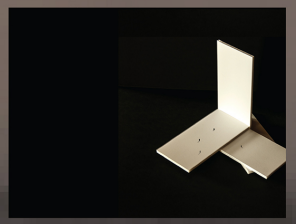
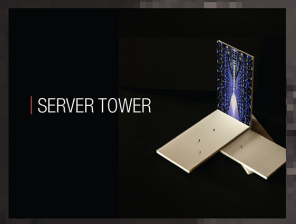
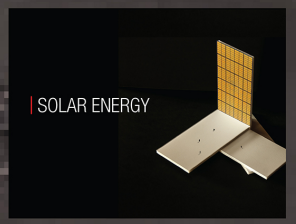
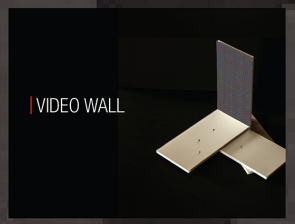
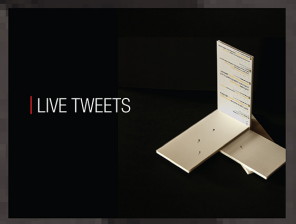
3) Consider the Implications of Keeping Your Archives "Dark" and Unavailable to Public Scrutiny.

4) Data Records Storage and Preservation Should Be Considered From the Beginning Stages of the Curation Process.

(Sources: Ess, 2012; CLAUSSIC, 2009; S. Li & N. Worby, personal communication, January 29, 2015; Small et al., 2012)

CURATE
Harvesting and Curation Issues

Here are some issues to keep in mind in the harvesting and curation of political Tweets:

- First, it is frequently difficult to find event-related hashtags without accidentally capturing a Tweet author's personal Twitter page.
- Archive-It can be set to block personal Twitter pages.
- An additional avenue is to search for thematic hashtags.
- However, Political Tweets Librarians may not be able to access every Tweet for a political event.
- One solution is to work alongside information specialists* directly involved with the political event for greater Tweet harvesting.
- *Journalists, news media broadcasters, political group leader/member.
- Secondly, there is more than one side to any political event and therefore, curation must be as unbiased as possible.
- To reduce bias, obtain as balanced collection as possible of Tweets from all sides of the political debate.
- This is a very time consuming effort, especially if the number of Political Tweets Librarians is limited.
- A possible solution is collaboration with other institutions devoted to political Tweets harvesting and curation.
- Collaboration lightens the load for everyone!

(Sources: Antracoli, et al., 2014; S. Li & N. Worby, personal communication, January 29, 2015)