ISSN 2324-7878 (online)
palrap.org

PaLA College & Research Division (CRD)

Pennsylvania Libraries: Research & Practice

Practice

Pennsylvania Perspectives of the 2016 Election

A Project to Collect Web and Social Media Content Around Significant Societal Events

Anthony T. Pinter, Ben Goldman, and Eric Novotny

Anthony T. Pinter is a PhD candidate in Information Science at the University of Colorado Boulder, Anthony.Pinter@colorado.edu
Ben Goldman is the Kalin Librarian for Technological Innovations at Penn State University Libraries, bmg17@psu.edu
Eric Novotny is History Librarian and Coordinator of the News and Microforms Library at Penn State University, ecn1@psu.edu

During the 2016 election, Pennsylvania was viewed as a crucial state not only for the presidential race, but also for a Senate seat, seats in the House of Representatives, and for state-specific positions. In response to the attention placed on Pennsylvania during the election, Penn State University Libraries undertook a project to document the discourse that occurred online. The resulting project, "Pennsylvania Perspectives on the 2016 U.S. Election," collected websites and Twitter data in order to document the people, voices, moments, and prominent issues in Pennsylvania. In this practice paper, we describe the project background, scope, collection methodology, lessons learned, and best practices that we discovered, in the hopes that it will inspire others to undertake similar projects to document important societal events at local, state, national, and international levels.

Introduction

In this practice paper, we describe creating "Pennsylvania Perspectives on the 2016 U.S. Election," a web and social media archiving effort aimed at documenting the people, voices, moments, and prominent issues in the Commonwealth of Pennsylvania during the final one hundred days of the 2016 election. The collection can be accessed via this finding aid (libraries.psu.edu/findingaids/9790.htm). In addition to exploring the project goals and design, this paper will discuss the method and scope of data collection, and share some lessons learned.

Throughout the 2016 campaign, Pennsylvania was described as a critical battleground state, with both candidates for President making frequent stops across the Commonwealth in the final months before Election Day. In the 2016 election cycle, Pennsylvanians voted on one of their two U.S. Senate seats, all 18 seats in the U.S. House of

Representatives, 25 of its 50 state Senate seats, all 203 of its state House of Representatives seats, and three key executive state-level offices (attorney general, treasurer, and auditor).

With so much attention and intellectual energy focused at the national level on the contentious race for President, our goal in the University Libraries was to document the Pennsylvania experience surrounding the election, by examining content found on websites and social media, especially those ephemeral viewpoints that often elude archives, and which can be easily lost to future scholars. What did it mean to be a "battleground state?" How were Pennsylvania's local elections influenced by national trends? What were organizations and individuals in Pennsylvania saying about the election as it unfolded? Where did groups and individuals stand on various important campaign issues? Our project sought to provide a window into such issues through examination of the online information and conversations reflected in Pennsylvania-specific websites and Twitter data.

Background

Penn State University Libraries began archiving websites in 2012, with a subscription to the Internet Archive's Archive-It service (archive-it.org), which provides a web-based dashboard for administering the capture, curation, and description of web content. The Internet Archive preserves the content collected and makes it publicly accessible via its Wayback Machine (archive.org/web) and partner-specific landing pages. Our initial institutional collecting efforts were aimed at enhancing existing curatorial goals. Our first collection, for instance, included over 100 websites in the psu.edu domain to support the mission of the University Archives. But as the program evolved, we have used the Archive-It service to explore emerging collecting areas, and collaborate with selectors across the Library, the University, and beyond.

Notably, web archiving has allowed us to explore ways to proactively collect current topics. Traditionally, archives and libraries acquire primary source material from individuals or organizations long after their creation; however, many genres of content traditionally collected by archives are now published on the web, and much of this content can be ephemeral. Many institutions are engaged in proactive collecting using Archive-It and other tools, especially aimed at government websites and election cycles in the United States and elsewhere in the world. Since 2008, the Internet Archive has collaborated with a variety of institutions to create an End of Term Web Archive (eotarchive.cdlib.org), a capture of all U.S. federal websites at the end of presidential terms. Archive-It partners worldwide have also created over 300 election-focused collections, which are accessible via this link (archive-it.org/explore).

In recent years, with the rise of social media, archivists and librarians have also begun developing strategies and tools for collecting content from these platforms. While content can be collected using Archive-It, there have been various efforts to capture content in ways that support computational research methods. With the popularity of Twitter as a central platform for public discourse and even political protest, particular emphasis has gone into developing tools that enable the capture of raw data from Twitter.

Many of the collections curated with such tools have focused on elections and social movements. Notable examples include a collection of four million tweets related to the 2015 Canadian Federal Election that was won by Justin Trudeau, and the more than 13 million tweets surrounding the events of Ferguson, Missouri, and the death of Michael Brown, who was shot by a police officer (Ruest & Milligan, 2016; Summers, 2014). In these and other instances, collecting is oriented around a keyword ("ferguson") or prominent hashtag ("#elxn42"). The researchers involved in both of these projects have analyzed the data collected and shared findings, demonstrating the potential research uses of Twitter archives.

Missing from the professional landscape is a consideration for how these disparate approaches — URL-based web archiving and keyword/hashtag based Twitter collecting — might complement each other if considered in tandem

when one is scoping a new project. Given the active effort by librarians and archivists to document campaigns, elections and public discourse at a national level, local content and dialogue is at risk of being excluded from the historical record. "Pennsylvania Perspectives on the 2016 Elections" was conceived as a way to fill this potential gap and capture the specific experiences of Pennsylvanians, as documented via these online platforms.

Designing the Project

From the initial planning phases, "Pennsylvania Perspectives on the 2016 U.S. Election" was conceived as a collaboration between librarians working in different areas of the University Libraries. Subject specialists in history, communications, government documents, and news/microforms were charged with setting the scope of the collecting effort, which included defining the overarching documentation strategy, identifying the specific sources of content to be captured, and the duration of the collecting activity. The digital archivist in the Special Collections Library performed the capture of websites, and a graduate assistant from the College of Information Sciences and Technology was tasked with collecting Twitter data. Subject specialists reached out to contacts in academic departments to gauge interest in the project and elicit feedback, which resulted in useful suggestions for content to add. The library subject specialists work closely with faculty and students in their disciplines, and are familiar with their research needs. It must be acknowledged, however, that having more direct contributions from political science, history, or communications faculty would have further enhanced the content selection process. Later in the project, students were enlisted to perform vital but laborious tasks, including identifying websites of candidates and political parties, and conducting quality assurance tasks on the websites collected.

Project scoping was accomplished in bi-weekly meetings of the subject specialists and the digital archivist over the course of several months. These meetings resulted in the identification of five overarching categories of content to be archived, with varying frequency, between the beginning of September 2016 and the middle of November 2016. Particular emphasis was placed on representing the diversity of political views found within the following categories:

- Election coverage of Pennsylvania-based media organizations from the top twenty media markets
- Websites and Twitter accounts of Pennsylvania-based political advocacy groups
- Facebook pages of university student political groups (e.g. the Penn State College Republicans or Villanova University College Democrats)
- Websites and Twitter accounts of Pennsylvania state political party websites
- Campaign websites and social media of Pennsylvania-based candidates for political office

In defining the scope, the subject specialists and digital archivist brainstormed and documented possible research questions that might be answered using the web archive. Some examples of these questions were:

- What did a particular advocacy group (e.g. Pennsylvania Chapter of the Sierra Club or the Philadelphia Tea Party Patriots) have to say about the election?
- How did these groups react to the results of the election?
- What was a particular group's (e.g. students, liberals, conservatives) perspective on the election?
- What was the response to (and impact of) specific notable events during the campaign?
- What was the influence of a campaign's message (e.g. how often was it shared/retweeted)?
- How aligned was the messaging of allied advocacy groups?
- How did coverage in traditional media compare to the most discussed topics in social media? Were there differences by region?
- What issues were most dominant in metropolitan areas like Philadelphia or Pittsburgh, compared to the rest of the state?

Once the scope was defined, the group set out to identify and inventory the specific websites, Facebook pages, and Twitter accounts to be captured. Group members received assignments, and identified websites using a variety of methods.

The subscription database *SRDS Media Solutions* was consulted to identify the largest news markets in Pennsylvania. Often used by advertisers, *SRDS* (Standard Rate & Data Services) provides information on newspaper circulation and media viewership for individual titles and by media market.

To identify student groups, we started with *U.S. News & World Report's* Colleges in Pennsylvania (www.usnews.com/best-colleges/pa). From that list, we chose 48 four-year public and private colleges and universities, with the goal of having a representative sample of private and public institutions of varying sizes and locations within the state. Selected schools range in enrollment from 690 undergraduates (Cheyney University of Pennsylvania) to 41,000 undergraduates (Penn State/University Park campus), and are located in all regions of the state. For the selected schools, we checked online for active student political chapters representing five main political parties with candidates in the presidential race: Democrat, Republican, Green, Libertarian, and Constitution (no Constitution or Green Party student chapters were located).

Selection of advocacy groups was open-ended, beginning with the identification of national chapters of familiar groups such as the National Rifle Association and the American Civil Liberties Union. Google searches yielded other groups, while social media connections uncovered many new groups. The linked nature of social media proved invaluable in identifying related groups. For instance, the Berks County Patriots yielded links to many allied groups, while in Philadelphia and Pittsburgh, chapters on the political left and right were often interconnected with groups of similar political persuasion. Each group was reviewed for evidence of recent electoral commentary. Omitted were groups with outdated websites, fewer than 100 followers on social media, or lack of political activity.

This initial selection process helped further narrow the scope of our efforts. A few news media websites were deselected due to perceived challenges in limiting the web capture to just election-related news stories. Advocacy group websites were selected based on recent political activity or commentary; if the most recent content found on an advocacy group's website was from 2015, for example, it was not included. Also, an immigrants' rights group that exclusively focused on social services was not included in our election archive. While national organizations such as Black Lives Matter or the Tea Party doubtless influenced the conversation in Pennsylvania, we settled on organizations with a distinct Pennsylvania web presence. Student groups were found to be primarily active on Facebook, yet many Facebook pages for student groups had not registered new posts in months or years, and so were not included.

The selection effort concluded with over 750 unique website URLs and over 300 Twitter accounts. In addition to the Twitter accounts, we identified seven hashtags related to Pennsylvania (including #pasen, #pahouse, #papolitics), and three that were not specific to Pennsylvania (#clinton, #trump, #election). The rationale for including these three additional hashtags was to collect a set of data related to commonly-used global hashtags that could be mined for keywords or phrases pertinent to Pennsylvania, or to explore which of the tweets with these hashtags might be geographically limited to Pennsylvania.

For planning and communication purposes, the group also developed a project overview document, which outlined the rationale for the project, scope, duration, and desired outcomes. For many of the partners on this project, web and social media archiving represented a new and exciting, yet unfamiliar, area of practice. We discussed in advance the technologies employed, the limitations of the intended tools, the preservation concerns, and the modes of discovery and access possible (including the limitations on access imposed by services like Twitter).

Creating the Archive

Archive-It provides its partners a browser-based administrative dashboard for creating new web archive collections, adding URLs to capture (referred to as "seeds" in web archiving nomenclature), initiating crawlers (software that visits and indexes documents found on identified websites), and then reviewing finished crawls for completeness and accuracy. Because the platforms, tools, and strategies used to publish content on the web are so varied, Archive-It recommends test crawling seeds as a kind of technical appraisal step. Test crawls are not only useful for identifying important content that was missed by a crawler, but also—since Archive-It subscriptions have yearly data budgets—in identifying the potential capture of unwanted data. The general workflow for creating a web archive is to test crawl a seed, review it, configure it, and then either test crawl again or perform a live crawl. A successful live crawl is typically followed by a quality assurance step.

Following this process, we identified a handful of problematic websites. One example was the election section of the Pennsylvania Cable Network (pcntv.com/election), which contained video broadcasts from campaign rallies across the state. The videos on this website were not being captured by a test crawl, and further investigation found the videos were delivered using a new technology for which Archive-It's engineers had not yet developed a solution. In another example, we ended up not crawling the website of the National Rifle Association, because it proved difficult to scope that website to just Pennsylvania-specific content.

While many websites use common web publishing platforms that make them easy to capture (campaign websites, for example), some websites can inadvertently capture more content than intended. In this regard, Facebook is one of the most challenging. With some careful configuration, Archive-It can be used to capture Facebook pages, but it is recommended that each seed be limited to one gigabyte per crawl. This presented a dilemma for the project team. We had identified nearly 500 Facebook pages, which we intended to crawl several times each over the course of the ten-week collecting period. At one gigabyte each, just crawling the selected Facebook pages would have exceeded our annual data budget of 750 gigabytes. This forced us to reconsider our documentation strategy related to Facebook. Our largest segment of Facebook pages came from Pennsylvania candidates for office, who were equally well represented by traditional campaign websites and Twitter accounts. By eliminating campaign Facebook pages, we reduced the number of Facebook sites to 122, resulting in a total of 470 websites overall. This decision gave us some flexibility with the amount of data to be captured, but the frequency with which we had intended to capture many of the websites would also have put us over budget. In the end, we decided to capture news websites every two weeks, but most every other site was captured once in early September and once just after Election Day in November. Whether because of irrelevancy of content or because of technical challenges, all deselected websites were carefully documented to help explain our decision-making to future users of the archive.

Our final crawls were launched the week after the election. This was followed by an effort to perform quality assurance on the crawls, which surfaced numerous missing pages. Once the missing pages were also crawled, we ended up with 448 gigabytes of data and 5.3 million documents from 471 websites. The finding aid for this web collection and the Twitter collection (discussed below) was previously mentioned in the Introduction. Further detail on how researchers can access the collection is detailed below.

Creating the Twitter Archive

The Twitter collection process can be broadly separated into two distinct phases: collection and management. We will report each of these phases individually. We will then discuss why collecting this data might be useful, including highlighting some existing use cases for Twitter data.

Collection

The cornerstone technology of our collection was twarc (github.com/edsu/twarc), an open-source suite of Python programming language scripts written by Ed Summers. twarc is run through the command line, and collects tweets in the form of *JSON* (JavaScript Object Notation) files, a human and machine-readable plain text format. Each tweet and its associated metadata are represented as a single line in the output JSON file. In addition to allowing users to collect tweets around a specific keyword, twarc also includes functions that allow for collection of tweets from specific users and around specific trends. Tweets can be collected as they are sent, or, through a search feature that collects tweets sent in the past seven to nine days.

We chose twarc as our collection tool after exploring similar tools, including Archive-It and Social Feed Manager (gwu-libraries.github.io/sfm-ui). twarc offered two notable advantages over other tools, making it an ideal fit for this project. First, it is open-source and has a low barrier for entry, because it only requires basic command line knowledge. Social Feed Manager is a more powerful tool, but would have required significantly more technical support. Second, JSON is a useful format that makes the archive immediately analyzable in several ways, which we felt was important for addressing the potential research questions we posed at the outset of the project. Archive-It can archive Twitter pages, but is not as readily usable for computational research. However, it is important to note that twarc is less effective than other tools in capturing embedded media. We discuss that caveat in the "Lessons Learned" section of this article, and address some examples of possible analyses in our "Next Steps" section.

We found that the best practice for generating robust datasets using twarc was to combine the tool's two collection mechanisms: collecting via streaming and running searches each week. This leads to duplication of tweets, but other projects have recommended this approach to avoid any potential gaps in collection. A deduplication script is also available within twarc's suite of utilities.

As previously discussed, our collection consisted of tweets from over 300 Pennsylvania-based accounts, ten hashtags related to Pennsylvania, and three global hashtags. We began collecting on September 6, 2016 and concluded on November 14, 2016. In total, we collected for just over 69 days, and searched the hashtags and accounts once a week (11 times in total). We collected 12,645,972 tweets, totaling 74.79 GB. In Figures 1 & 2, we display the timelines for the two datasets, highlighting both the size of the datasets and a potential way of visually displaying this data.

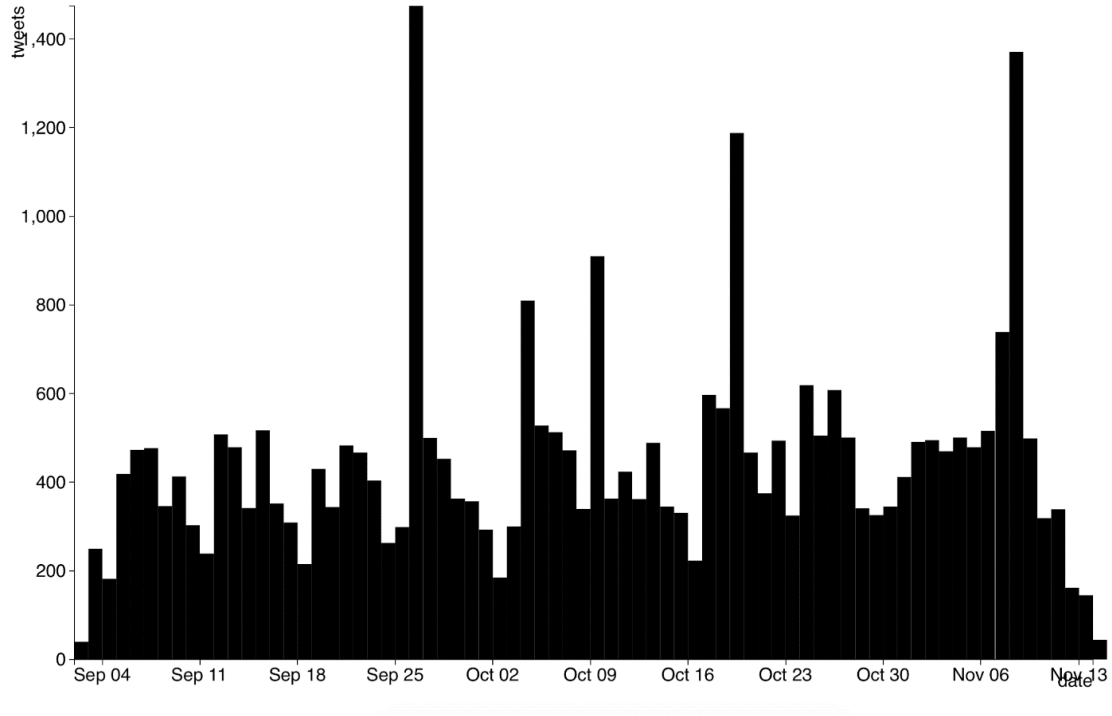

Figure 1

PA-based Dataset Timeline

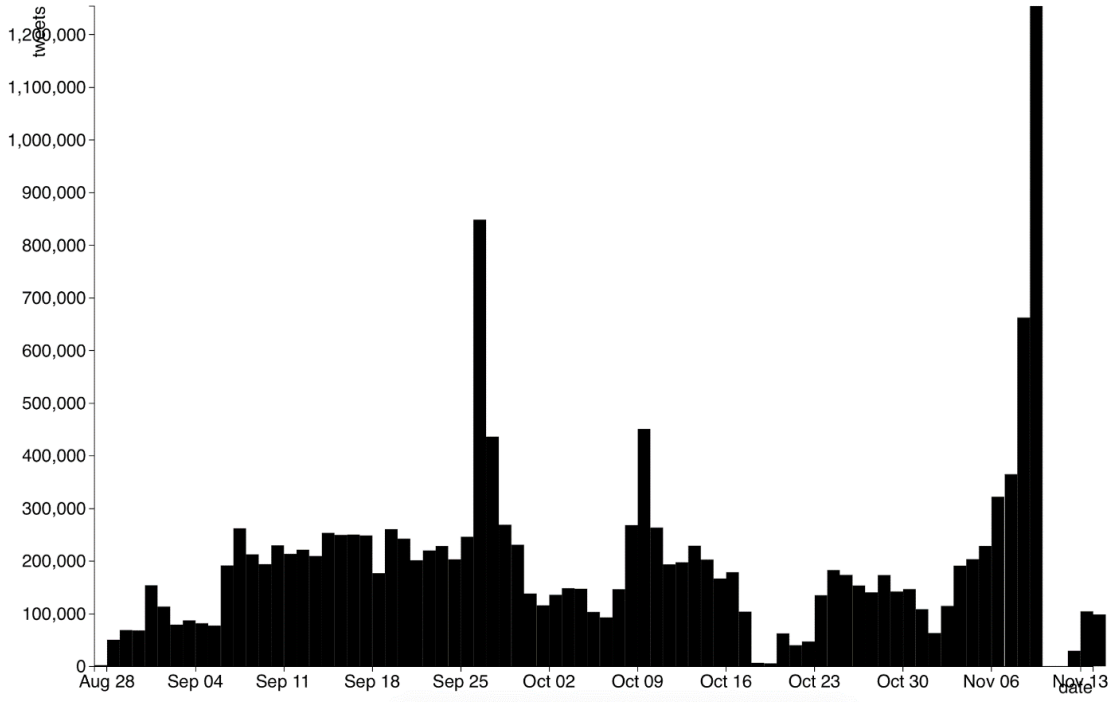

Figure 2

General Election Dataset Timeline

Management

With such a large dataset, it comes as no surprise that there are difficulties to contend with when considering how to manage the data. Many of the difficulties stem directly from the size of the set itself; at scale, aggregating tweets into a single file can be difficult to open using certain common text-viewing applications. Thus, using the command line is the best way to manage large datasets. There were several processes that required commands or scripts to automate.

First, the *cat* command has a variety of functions that can be used to measure and manage large datasets. The cat command can be used to measure the number of lines in a file (in this case, analogous to the number of tweets, as one tweet equals one line). It can also be used to combine multiple files into a single file. The latter function is useful if a collection is made of several different files (as our weekly collections were), or if the dataset is split into smaller files for ease of transmission and storage.

Splitting the files into smaller files is particularly sensible when a specific sub-dataset is too large to store. For example, our collection using the hashtag "#trump" was too large to easily share with project partners, so we utilized the 'split' command to parse the collection into equally-sized chunks.

It is also important to note that twarc includes a variety of scripts that are useful in managing collections generated using twarc. Within twarc's utilities folder, there are scripts that filter tweets by author gender or by the presence or absence of geolocation tags (e.g., coordinates or place values). There are additional scripts that can remove duplicate tweets, sort by tweet id (analogous to sorting by time), remove tweets that were published before a specified date, remove retweets, and see the clients that users were using to send tweets. These scripts are useful in addressing research questions such as the ones discussed above, in the "Designing the Project" section.

However, the most important data management function in twarc is the ability to hydrate/dehydrate tweet files. Twitter's terms of service set strict guidelines on sharing datasets of tweets. Every tweet has a unique identifier (ID), and those identifiers may be published so that researchers can reconstitute the data on their own (a process known as 'hydration'). The twarc 'dehydrate' command takes a JSON file of tweets and outputs a text file of just the tweet IDs from the JSON file. Conversely, the twarc 'hydrate' command takes a text file of tweet IDs and outputs a JSON file with full tweets. These commands are useful not only for making datasets widely available, but also for supplementing new collections with existing collections to create richer data around societal events.

Lastly, we packaged JSON files as *bags* using the BagIt protocol (tools.ietf.org), which offered a modest layer of preservation around the data by generating checksums and file manifests. This proved particularly helpful for working with the data, which often required transmitting the data between computer environments to perform some of the activities mentioned above. For instance, when we split a set into smaller pieces and wanted to include information about how to reconstitute the set into one workable file, bags allow for easy categorization of the data they contain, and can be configured to include information about the origin of the data encapsulated by the bag. The final dataset was stored as bags alongside our other born-digital collections in the Special Collections Library.

Discovery and Access

With capture of websites and Twitter data completed, we set out to enable discovery and access of the collection. The archived websites were made public and are now accessible online via at Penn State's partner page in Archive-It (archive-it.org/collections/7694). A collection description was added, in addition to minimal seed-level metadata (Title and Date). Seeds were also organized into groups according to the categories identified earlier in the project.

Twitter's API terms of service, as interpreted by most librarians and archivists, imposes certain limits on access. Unlike Archive-It, we cannot make the dataset readily accessible online; though we can provide offline access to any Penn State-affiliated researchers. To enable use by non-Penn State researchers, we published the lists of tweet identifiers to the institutional repository Scholarsphere, (scholarsphere.psu.edu/collections/cz30ps790), along with an overview of the limitations on access and tools that can be used for recreating the data. Efforts to reconstitute the data, it should be noted, could result in some data loss, as tweets are deleted or made private by account holders.

These nuances of access have also been described in the archival finding aid mentioned in the introduction of this paper. The finding aid provides both an overview of the collection and a detailed inventory of the archived websites and Twitter content. This collection is discoverable alongside all our other archival collections, via the Penn State Libraries' website. Going forward, we plan to promote the archive more broadly to the library and scholarly community through groups such as h-pennsylvania (networks.h-net.org), which is an e-mail list for scholars interested in the history and culture of PA, and also via the Pennsylvania Library Association.

Lessons Learned

Despite conceiving of the project as both a website and Twitter archiving endeavor, the steps involved were quite divergent, and presented different challenges. While our subscription to Archive-It has lowered the technical barriers to capturing websites, a Twitter archive was an entirely new and unfamiliar technical challenge. When successfully captured, Twitter data is entirely predictable. In contrast, the results of crawling any individual website are never predictable. Some websites require several tests, some crawl very easily, and some not at all.

Communicating the unpredictability of web crawling turned out to be a necessity and challenge. The library selectors in History, Communications, Government, and News/Microforms, who enthusiastically embraced the project and set the scope, were quite naturally disappointed that some selections had to be reconsidered due to size and technical challenges. Our discussions around these points highlighted some opportunities for setting better expectations up front. As our web archiving efforts expand to include more collaborators, it will be important for us to create some documentation around the process that will help all involved understand the challenges from the outset of the project.

In terms of the number of seeds selected, this project ended up being our largest curated web archive collection. It should not have come as a surprise that as we engaged more selectors in this process, we more quickly used our allotted data budget. Nevertheless, our budget forced us to reduce the frequency of captures, which potentially created a more limited historical record. In the future, we would recommend that projects attempting to comprehensively capture websites around a specific known event begin planning much sooner. We might also have been able to alleviate some of these data limitations by exploring other web archiving technologies (such as WebRecorder for Facebook pages, or YouTube-dl for videos).

Archiving websites using Archive-It primarily requires resources and training, but Twitter archiving requires more technical competency. Tools like Social Feed Manager would have demanded significant technical support for us to implement. twarc, by contrast, only required some familiarity with the command line and Python. The data we collected from the Twitter streaming API essentially ran non-stop from a dedicated computer for the duration of the project. In retrospect, it would have been simpler to establish a virtual machine and stream Twitter data from there.

While we understood that our use of twarc would prevent us from capturing media embedded in tweets, we had not considered the possibility of capturing media and images using other tools. Had we looked closer at the tweets as we collected them, we might have undertaken an effort to separately capture or sample such content. In general, it would be useful to do some analysis of Twitter data at intervals throughout the course of such a project to identify any trends in the content that might suggest adjustments or additional steps to enhance the archive.

Next Steps

Moving forward, we plan to explore technologies that would support more advanced research options for the websites collected. Archive-It provides some advanced tools for examining metadata elements in web pages, defining which pages across the collection link to one another, and identifying named entities across the collection. Here is a link to more information about Archiv-It Research Services (webarchive.jira.com/wiki/discover). While it's not necessary for archivists and librarians to perform this level of research, we feel that we'll be better able to support potential research use cases with a better understanding of these tools and their capabilities.

As discussed above, the Twitter data was captured in a way that supports computational analysis, and we have already begun working with the dataset. A single tweet consists of a combination of thirty-five different fields that identify specific characteristics of that specific tweet. These fields include sub-fields such as tweet ID, tweet text, and user information. The wealth of data present in a single tweet means that large datasets among tweets can be analyzed in different ways to provide insight about how the Twitter public encapsulated in a dataset felt about specific issues.

Returning to our sample research questions as examples, we outline how each question might be approached using the metadata contained in tweets in Table 1.

Table 1

Original Research Questions and Potential Means of Answering Them

Original Question	**Potential Method of Answering**
What did a particular advocacy group (e.g. the Pennsylvania Chapter of the Sierra Club or the Philadelphia Tea Party) have to say about the election?	Separate that account out of the dataset using the 'user' sub-field, then pull the text out of the 'text' field for textual analysis.
How did these groups react to the results of the election?	Separate that account out of the dataset using the 'user' sub-field, then pull tweets that were published post-election day, then pull the text out of the 'text' field for textual analysis.
I want to examine a group perspective on the election (e.g., students, liberals, conservatives).	Separate those specific groups' accounts from the main dataset, then pull the text out of the 'text' field for textual analysis.
I want to analyze the response to, and impact of, specific events.	Pull tweets from specific days, then pull the text out of the 'text' field for textual analysis.
I want to measure the influence of a campaign's message (how often was it shared/retweeted?)	Use the 'retweeted' and 'favorited' sub-fields to quantify the reach of a specific tweet.
I want to know how similar allied groups were in their messaging.	Compare the tweets between two (or more) groups by separating them from the main dataset and then pulling the text out of the 'text' field.
How did coverage in traditional media compare to the most discussed topics in social media? Were there differences by region?	Compare the Twitter public's perceptions of events to traditional media coverage by comparing tweet text with news articles in a sentiment analysis.
What issues were most dominant in metropolitan areas like Philadelphia or Pittsburgh, compared to the rest of the state?	Use the 'geo-location' (if available) or the 'place' sub-field to create a dataset of tweets from a specific area, and then compare the tweet texts between these geographic datasets.

These questions and answers are far from exhaustive, and we only suggest a single approach to answering each of them. There are multiple ways to approach an individual question, and the amount of data associated with a single tweet means that there are limitless potential avenues of analysis. Thus, the potential for meaningful research to spring from a Twitter archive provides a strong argument for collecting and publicizing this type of archive to a wide audience.

Finally, we hope to develop a framework or specific implementation how-to and "checklist" to assist other librarians, archivists, and researchers replicate the approaches we took as part of this project. Creating this will enable others to avoid some of the pitfalls we encountered, encouraging more active collecting around societal events, and broadening the potential of this media as a significant archive and research source.

Conclusion

Despite some limitations and compromises, the collection team (subject specialists and digital archivist) remains enthusiastic about this proactive and collaborative approach to collection building. Particularly exciting is the ability to capture *engagement* and the voices of individual Pennsylvanians. Social media offers the opportunity to extend archives beyond a traditional focus on institutions and prominent figures. Those perspectives are already well represented in the historical record, but our project provides researchers with access to a broader historical record. Collection of the Twitter and Facebook content reveals not only what major actors had to say about the election, but also how individuals of all persuasions responded, through likes and comments and retweets. In capturing this specific historical moment, we hope to enable future scholars to ask new questions, and explore the 2016 campaign in all its complexities.

References

Ruest, N. & Milligan, I. (2016, April). An open-source strategy for documenting events: The case study of the 42nd Canadian federal election on Twitter. *Code4Lib Journal* (32). Retrieved from http://journal.code4lib.org/articles/11358

Summers, E. (2014, August 30). A Ferguson Twitter archive. Retrieved from https://inkdroid.org/2014/08/30/a-ferguson-twitter-archive