SPECIAL SECTION: ARCHIVES AND TWITTER

Preserving the Voices of Revolution: Examining the Creation and Preservation of a Subject-Centered Collection of Tweets from the Eighteen Days in Egypt

Timothy Arnold and Walker Sampson

ABSTRACT

In 2011, Hosni Mubarak abdicated his position of president-for-life after peaceful protests across Egypt. Demonstrators in these protests used social media platforms like Twitter to communicate directly with a global audience, but tweets are ephemeral and there are no standards or best practices for their collection and preservation. Using the revolution in Egypt as a case study, this paper serves as a guide to collection developers who are interested in collecting subject-centered collections of tweets. We will discuss how to collect tweets using Twitter's application programming interface (API) as well as collection development issues related to Twitter's role in the Egyptian revolution. These issues include determining the scope of the collection, quantitative and qualitative collection methods, separating signal from noise, and navigating vernacular and formal languages. We will consider the Twitter platform from an archival perspective and discuss best practice in archiving a collection of tweets. To present the kinds of materials that a subject-centered collection of tweets might include, we will conclude with a brief examination of tweets collected during the eighteen-day revolution in Egypt.

KEY WORDS

Twitter, Tweets, Social Media, Arab Spring, Egypt, API, Born-Digital, Digital Curation, Digital Collections, Subject-Centered Collections

On December 17, 2010, in a small town named Sidi Bouzid in Tunisia, Mohamed Bouazizi lit himself on fire in front of the governor's office. Bouazizi was a street vendor. His father had died, and as the eldest child in a large family, many of the responsibilities of providing for his family fell on him. He had graduated from university with a bachelor's degree in computer science, but like many in his generation he was unable to find a job and the financial need of his family drew him to informal labor.

The story of Bouazizi is typical of young people across the Middle East who are finding that, although they dedicated many years of their lives to education and career building, they have little or no opportunity to reap the rewards of their labor. Epidemics of unemployment among the young in the Middle East and North Africa (MENA) have led them into fields of work that are either unsanctioned or dubiously sanctioned by state government. Consequently, the battles between youth in the MENA region and the state police have become a flashpoint and a symbol of the greater struggle that youths face for basic financial solvency.

Whereas Mohamed Bouazizi's self-immolation became a symbol for the struggle of youth in Tunisia, another young man in Egypt named Khaled Said became a symbol of state violence. Said was beaten to death by Egyptian police outside a café in Alexandria while a number of bystanders looked on. Photographs of Said's beaten body were posted on Facebook and went viral. Like Bouazizi, Said's murder became a rallying cry for youths who were tired of constant police harassment, lack of opportunity, cronyism, and surveillance. Wael Ghonim, the creator of Khaled Said's memorial Facebook page and a young executive at Google, sent out a request to the page's subscribers to join a rally at Tahrir Square in Cairo on January 25, 2011—the "Day of the Police" in Egypt, a day to give thanks to the police for the alleged protection that they provide to Egypt's citizens. While the demonstrations were intended to draw attention to police brutality, their scope grew very rapidly. Eighteen days after the first gathering of protestors at Tahrir Square, Hosni Mubarak, who had ruled Egypt for over thirty years, peacefully abdicated his position of president-for-life, and the transition from an authoritarian to a democratic regime started to develop.

The role of social media in what has become known as the "Arab Spring" has received much attention. In the Western press, the revolutions in Tunisia and Egypt are often dubbed "Twitter Revolutions" or "Facebook Revolutions."[1] Others in the mainstream Western media have attempted to prove that social media had only a dubious relationship to the efficacy of social movements.[2] The term "Twitter Revolution" implies an underlying assumption that the use of Twitter by young people in Egypt and Tunisia afforded a kind of power that they did not have prior to the introduction of social media to the region. On the other hand, some have pointed out that identifying social media as a monocausal factor in the development of the Arab Spring disempowers the

actual actors who brought about the sweeping political changes of January and February 2011.[3] Meanwhile, others have said that social media played only a minor role in the organization of the uprisings and that more traditional forms of media like satellite television—particularly the Qatari international news network, Al-Jazeera—were the main cause of the Arab Spring.[4]

In September 2014, a conclusive explication of the role of social media in the Arab Spring remains forthcoming. It seems unlikely that researchers will be able to determine the role that social media play in events like the Arab Spring without access to the actual historical records of online protest activity. While many researchers have written about the role of social media in the Arab Spring, few have conducted rigorous content analysis simply because the content is unavailable to scholars who do not have the ability to build tools to collect data from Twitter. As Laila Shereen Sakr noted, "A review of scholarship on media and the Middle East reveals a lack of engagement with digital media content, whether as primary sources or in critically questioning the tools and analytics provided."[5]

The archivist's role is to collect, preserve, and provide access to historical records for researchers. However, it has also been observed that "the rate at which digital preservation capabilities develop does not match the rapid rate at which human rights actors produce born-digital documentation."[6] We hope that this article will start a dialogue about how to collect and preserve subject-centered collections of tweets so that researchers can better understand Twitter and its effect upon on-the-ground activity in events like the revolution in Egypt.

The article is divided into four sections. In the first section, we explain what Twitter is, how "tweeps" (Twitter users) employ the platform, and how developers of Twitter have created functionality to support use that emerged organically from Twitter users. In the second section, we discuss collection development. We begin this examination by introducing three of the most common application programming interface (API) commands researchers can use to collect tweets. We then discuss some collection development issues related to Twitter's role in the revolution in Egypt. Some of these issues, such as the growing use of nonstandardized Arabic dialects in communication via Twitter, are unique to the MENA region; other issues, such as the Egyptian government's use of Twitter as a vehicle to spread misinformation, are not unique. In the third section, we detail long-term preservation concerns for this type of collection and methods to ensure that such collections remain available in perpetuity. Finally, in an appendix, we offer some examples of the types of tweets that one could collect from an event like the revolution in Egypt to illustrate the kinds of materials that researchers might want to access from such a collection.

An Introduction to Twitter

Social media come in many forms. Facebook is a "social-networking service," LinkedIn is a "professional-networking service," and Reddit is a "meme-generator." Twitter is a "micro-blogging application."[7] Microblogging "allows users to exchange small elements of content such as short sentences, individual images, or video links." Some social historians relate Twitter to the practice of recording daily activities in diaries.[8] "Twitter can be compared to earlier sources of personal information . . . many of the earlier sources contain mundane or trivial pieces of information that, in aggregate, can tell a detailed and authentic story about everyday life that is difficult to find elsewhere."[9] Though many wonder why a medium on which users discuss what they ate for breakfast has become so popular, Twitter can be used for much more than logging mundane activities and phatic communication. The type of use most relevant to the subject of this study is event following. Twitter users educate themselves about events as they unfold through interaction with other Twitter users, some of whom are active in the actual events.[10]

Unlike other social media platforms, most notably Facebook, Twitter provides relatively few constraints in terms of the way the platform can be used. Users are provided with essentially an empty canvas which they can use in any way they see fit. The only constraint is that the message must be 140 characters or fewer. Any user can quickly and easily establish connections with any other user anywhere on the globe. Twitter etiquette does not prohibit users from "following" other users with whom they have no preexisting "IRL" (in real life)

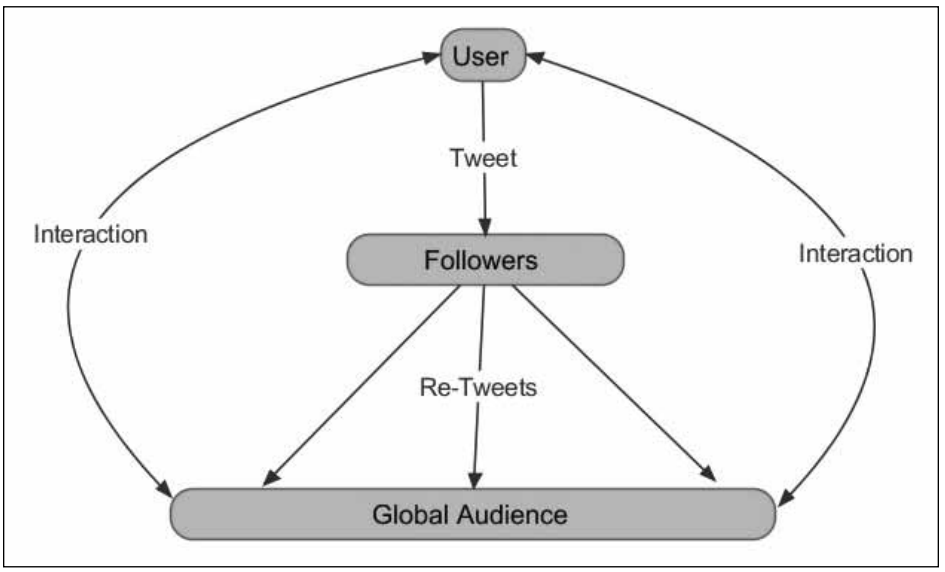

FIGURE 1. As Twitter users tweet and retweet, they spread information globally.

relationship. Messages created on Twitter in the form of "tweets" can originate from a single user and reach a global audience through the user's followers. Figure 1 illustrates how information is spread via Twitter.

As Figure 1 suggests, information circulates through complex, many-to-many patterns. Members of the "global audience" who receive the information in the form of the original user's tweet can respond to and interact with that user.

The many-to-many interaction on Twitter is facilitated through a set of symbols and functions that evolved organically. Before developing functionality, Twitter developers examine the ways in which Twitter users come to agree upon forms of use. Twitter users employ certain symbols as a shorthand to indicate actions as demonstrated in Table 1.

Table 1. List of Twitter Symbols

Symbol	Name	Use	Example
@	Mention	References another Twitter user.	"Hanging out with @ timmerino tonight."
#	Hashtag	Indicates the subject of the tweet.	"Parliament is still drafting the constitution. #Egypt"
RT	Retweet	Indicates a quote from another Twitter user, typically followed by an @ mention.	"RT @timmerino I didn't like that movie very much. #Skyfall"
MT	Modified Tweet	Indicates a paraphrased retweet.	"MT @timmerino I didn't like #Skyfall."

The "@" symbol is used to indicate a reference to another Twitter user. When Twitter developers noticed that users were employing this symbol, they developed a function whereby the @ symbol and username would automatically generate a hyperlink to the mentioned user's account page. Another function was developed whereby the user could see any instance in which she or he was mentioned using the @ symbol, thus facilitating interaction.

The "#" symbol, or "hashtag," preceding one or a group of words is used as a subject tag that helps users search for subjects with multiple words. "#Jan25," for example, was one of the most popular hashtags used during the revolution in Egypt.[11] Twitter developers have introduced a number of functions related to hashtags. Selecting a hashtag will allow the user to see everything that other users are saying about the subject that the hashtag indicates. Twitter also tracks hashtags and displays them on a list of "trending topics" so that users can see what subjects are being discussed. Both of these functions, and a number of others, are emergent in the sense that the developers did not produce them when Twitter was created. The functions emerged from communication among Twitter users and were adopted broadly across the platform before the developers

FIGURE 2. Twitter's interface has been refined based on the ways users tweet.

created additional ways to use the functions.[12] Figure 2 displays some of Twitter's current functionality.

The box on the bottom left, "United States Trends," displays the hashtags most popular in the United States at the time this screenshot was taken. In the main "Home" section, note the hyperlinks prefaced by the @ and # symbols. The "Notifications" tab at the top left of the image leads to a page where the user can see all of the tweets in which his or her Twitter handle is mentioned.

Collection Development Strategies

Twitter's Application Programming Interfaces (APIs)

In this section, we discuss tweet aggregation. Twitter has two APIs through which one can collect data related to either search terms or users. An API is a method through which one piece of software can communicate with another piece of software. In this case, we are using the API to have a simple client (our software) request data from the Twitter platform (the other piece of software). The various commands within the Twitter APIs return different kinds of data and have different limitations. The API commands we discuss in this section are current at the time of this writing. Note that the APIs and their functionality

are not static features. These APIs are developed and maintained by Twitter and are subject to change.

REST API, the *Search* Command

Twitter's first suite of API calls are collected in their REST API. REST stands for representational state transfer and is a common architectural style for World Wide Web services and protocols, including the ubiquitous HTTP (hypertext transfer protocol). Within this group, the Search API call allows aggregation of tweets that include a specific search term sent to Twitter's servers within the week prior to the query. Note that the default number of requests allowed for a standard search API call is 180 search requests per fifteen minutes. One search request will yield a maximum of 100 posts, so for a standard search API request, one should anticipate a maximum of 180 requests times 100 posts resulting in 18,000 tweet returns per fifteen minutes. This rate is probably fine, unless one is tracking a globally significant event as it occurs, in which case the Stream API would be more effective.

REST API, *User Timeline* Command

With a list of tweeps who have used a search term, one can request user data through the User Timeline call, also part of the REST API suite. The types of user data one can collect from the User Timeline request include

- Username
- Full name (if listed)
- Location (if listed)
- Bio (if listed)
- Tweeps the user is following
- Tweeps following the user
- The entire list of the user's tweets up to 3,200 tweets

So, although one can only collect tweets from the prior week using the Search call, that information can be used to determine which users are discussing one's topic of interest, gather the usernames from those tweets, and then collect the last 3,200 tweets from that user's Twitter profile page. While many Twitter users may not have 3,200 tweets, activists who were engaged in the revolution in Egypt often tweeted more than that. Note that if a user chooses to protect his or her tweets, none of the data from the user's Twitter page, including the actual tweets, can be collected using any of the APIs.

Stream API

The Stream API was released in 2010 and allows the user to collect tweets with a designated search term the moment the tweets are sent to Twitter's server. The potential for collecting tweets using the Stream API is vast, as Twitter yields up to 1 percent of its tweet volume to each individual Stream API request. As millions of tweets are generated every hour, the Stream API affords great potential for collecting tweets that include a search term. Bear in mind that the Stream API cannot be used to collect tweets sent to Twitter's server before a query is sent to the API. It only captures tweets as they are generated. In the next section, we discuss the role of these activists in the revolution and why one might want to structure a collection around their tweets.

Determining Scope

The first step in building a search is to decide whether to collect tweets from a specific user or set of users or to collect more broadly. The scope of a collection can be limited to users who indicate that they are tweeting from a certain time zone or who indicate they are tweeting from a specific location. A search might also be limited to those users who play the role of opinion leaders or organizers. Identifying this group may take time. By far the best way to do so is to become embedded in a community that discusses one's subject of interest, watch who is retweeted most frequently, and try to identify which individuals are being mentioned the most. Who is included in and who is excluded from conversations? Once two or three opinion leaders have been identified, finding more becomes easier because they know each other, form relationships with each other that are mediated through Twitter, and often exclude nonopinion leaders from their discourse. Once the users whose tweets will be included in a collection have been identified, the User Timeline API call can be used to collect their user data and their most recent 3,200 tweets.

This qualitative approach is, in our opinion, the best way to create a subject-centered collection of tweets based on the tweets of opinion leaders. For a more mixed-methods approach that is less subject to bias,

- Enter the desired search terms into either the Search API call or the Stream API;
- Enter the tweet data into statistical analysis software such as Stata;
- Decide which criteria will be used to determine how to identify the group of users. The options include
    - The user with the most tweets;
    - The user who is mentioned the most; or
    - The user who is retweeted the most.

- Use the User Timeline API call to collect the data from the users identified in the above steps.

No matter which method is used to determine the scope of a collection, temporality is always a key concern. Though the importance of collecting early will vary, querying Twitter's APIs as early as possible will help ensure a more comprehensive collection, particularly in collecting tweets from a globally significant event as it unfolds.

Identifying Search Terms

While the above guidelines may seem sufficient for most projects, certain hashtags are used more than others, and, during a globally significant event, the volume of tweets may be such that a complete collection of tweets with a certain hashtag is impossible to compile. An incomplete set of tweets that includes one's search term may be fine. To collect a complete set of tweet data that includes a search term within a specific period of time, other hashtags being used alongside the more popular hashtags can be monitored. For example, the tweet "We are marching to #Tahrir square. #Jan25 #Mubarak #7uriyya" contains four hashtags: #Tahrir, #Jan25, #Mubarak, and #7uriyya. The first three hashtags, and particularly #Jan25 and #Mubarak, were used very commonly to indicate that a tweet was referring to activities or statements related to the demonstrations in Cairo during the eighteen-day revolution. #7uriyya, however, is a transliteration of the Arabic word for "freedom," and, because it could only be understood by Arabic speakers who are also able to use Latin characters to transliterate Arabic words, it was used much less commonly and by a very specific set of users who had at least some knowledge of both Arabic and English. So, finding the less commonly used hashtags might be a worthwhile strategy to identify a niche set of users.

Another advantage of identifying hashtags both used by and pertaining to specific demographics is that using them in queries collects less noise. Spambots will often use popular hashtags to draw the attention of Twitter users to a website. If spam is not within the scope of a collection, querying less commonly used hashtags will increase the chance of collecting only tweets that are relevant to a specific subject. Websites like twitblock.org have begun compiling lists of verified spammers on Twitter.

Vernacular vs. Formal Language

As noted above, #7urriyya is a transliteration of an Arabic word in Latin script. Vernacular language is an important consideration in determining which search terms to use. Spellings of vernacular Arabic words are becoming codified

for the first time since the introduction of social media to the region. Until very recently, vernacular Arabic was not written at all; formal Modern Standard Arabic was the only language used for written communication. A search word in formal Arabic may be spelled differently from its colloquial equivalent, and, at times, the search term might be transliterated using Latin characters. In many instances, the word might be entirely different. For example, the colloquial word for "today" in the Cairo time zone is *a-nahar-da*, and the formal Arabic *al-youm* is rarely used. *Al-youm*, therefore, would likely not make an effective search word, even in its varied spellings.

Tweet Veracity

One final and important collection development issue regards the veracity of data. Twitter affords a degree of anonymity which is very useful for on-the-ground activists who want to remain clandestine, but also allows a space where users can misrepresent both themselves and the events they discuss. While they arrived to Twitter somewhat late, repressive governments have found ways to use it to spread misinformation and confusion.[13] This complicates determining which tweets are legitimate representations of on-the-ground events and which are not. One way to get around the issue of veracity is to focus the collection on a group of opinion leaders as previously discussed. While this is not a fail-safe method, it will certainly filter out most of the noise from both repressive governments and spambots if such tweets are outside the scope of a collection.

As a last point, we note that the veracity of any tweet should not be taken for granted. Unfounded rumors can spread as easily via Twitter as they do in nondigital forms of communication; however, this should not affect the value of a subject-centered collection. A researcher might be interested in the generation and dissemination of rumors, and network analysis of a subject-centered collection of tweets might be the perfect way to conduct such a study.

Collected Tweets as an Archives Object

In the previous sections, we described Twitter use in a particular global event and our collection technique for this set of tweets. We would like to turn to preservation of the dataset. Before examining the specifics of this preservation work, it is worth briefly reflecting on the general characteristics of Twitter as a platform and tweets as archived objects.

Twitter Data as a Web Archives

Broadly, we find it accurate to characterize a tweet collection as a Web archives as described in the "Web-Archiving" report to the Digital Preservation Coalition: content captured from the live Web for future use or retention in an archival setting.[14] Specifically, a tweet collection falls under the rubric of user-generated Web 2.0 material. Within this emerging field of archiving practice, tweets afford significant advantages compared to general Web content such as websites and pages, while still sharing some of their drawbacks. Similar to collectors of Web content, researchers must contend with the complication of linked content. A tweet can link to images, video, websites, and other media. Very often such content is the subject of the tweet, increasing the importance of the linked content to the post's meaning. In addition, URLs for such content are usually mediated by a URL-shortening platform such as TinyURL, bitly, or Twitter's own service, t.co, among many others. This further complicates retention. The 301works.org project, which allows URL shortening services to provide backups of their URL mappings, provides a small, but not absolute, hedge against the loss of linked content.[15]

As is true for Web content, tweets are not categorized for users by the underlying platform. A Twitter research project therefore confronts the same complication as a subject-centered Web archives project: researchers must determine the scope of the capture, as well as the means of achieving that scope through search terms, users, and so on. While a Web archives is compelled to archive a site repeatedly to convey and capture all the dynamic content it contains, a Twitter collection must contend with the rate limits imposed by the API, along with constraints in network latency, which may prevent some tweets from being returned.

However, tweets share consistent features as part of a coherent microblogging network, perhaps more so than other social networking platforms. Twitter more strictly enforces permanency in the posting event. Tweets may not be edited or updated. Although users can remove a tweet, this action is not central throughout the platform; rather, the tweet is removed from the user's timeline. This prevents further views but does not remove retweets or previous captures of the content by other parties. There is no need to reharvest the same resource over time. This makes a Twitter collection more suited to "event-based" collecting than is other Web content.[16]

In addition, the visual or aesthetic aspect of a tweet is usually of minimal concern. Therefore, archivists often need not be bothered with preserving design aspects of a tweet; it is rather the "content" in the traditional sense of the term that is of interest. There are exceptions: tweeps may use extended

characters and white space to create an aesthetic statement. However, we find this to be an exceptional use of the platform.

These common characteristics of tweets may lead one to assume that a spreadsheet of captured tweets is an apt representation of their content, but such a representation will lack key features and context of the tweets. Similar to a Web archives, which can be described as not simply a copy but as a "double construction" made first of the selected URLs to crawl and second the specific versions of those URLs, a Twitter archives is a "reborn, unique and deficient" version of the original online material.[17] The archived set of tweets is quite different from the actual tweets as they were made online, as the former represent the scoping decisions of the researchers and the particular affordances of their tools. Perhaps more so than in a Web archives, significant experiential and contextual information is subject to loss. Gone are profile pages, user avatar images, formatting, and the unique sequence of tweets any user would see in his or her timeline, while conversations and linked content may be partially or entirely lost.

The Importance of Documenting Methodology

As Stine Lomborg has stated,[18] a tweet collection is akin both to field data and to a collection of textual documents. In terms of preservation, this means that a record of the methodology used in collecting the tweets is of paramount importance along with the tweets themselves. Without this record, the dataset loses vital context and future value.

We define *methodology* here as the tools and commands, as well as the decisions behind their use, that generated the collected tweets. A record of methodology guards against many of the hazards found in large datasets and within the emerging interest in "big data." Here we define *big data* as a set of data large enough that tool-based analytical techniques are needed to derive meaning. In this case, it would not be possible to simply read through the collected tweets and reach a meaningful conclusion. Hazards of big data include assumptions of objectivity, accuracy, and the equal value of each data point.[19] Records of how tweets were gathered, by whom, at what time, and with what tools can help address these assumptions and articulate incompleteness.[20]

Moreover, we believe the specific methodology used to collect tweets will provide further insight to future researchers on the Twitter platform and the content hosted there. As Niels Brügger and Niels Ole Finnemann noted, future humanities work can benefit from modifying "the uniform concept of the computer and 'computing' to a concept of digital media reflecting the fact that the functional architecture of computers is variable and subject to ever ongoing developments in different directions."[21] Documentation of methodology builds

toward understanding the functional architecture of Twitter and disabuses notions of the computer as a pure computation or mathematical platform.

Finally, a record of methodology assists in de-anonymizing the mined set of tweets by identifying the researchers and tools involved. It also provides a space to take credit for the techniques used to gather the tweets, which in many cases will not be trivial. In short, "we need to place data generation on more of an equal footing with final outputs; to think of it in terms of authorship."[22]

Understanding Twitter as a Historical Object

In considering preservation of Twitter data, it is vital to understand Twitter as a changing platform. As described by Richard Rogers, the perception and arguably popular use of the platform has changed considerably over time, from a personal status messaging system, to a news medium, to a massive database.[23] Even more important for our purposes, the specific functionality and terms of use for the platform are subject to change at any time by the company itself.

As an example of this change, we can consider the fluctuation of the API. Although it is perhaps common to consider Twitter and its API as a single entity, each is a separate enterprise. Twitter had no API for many months after its debut in 2006 and no search at all until the purchase of the Summize company in 2008.[24] In early 2011, Twitter announced it would no longer "whitelist" selected apps, which allowed those programs larger tweet yields than others by granting a higher number of requests per hour.[25] Individuals as well could apply for these whitelisting privileges. Outside of major changes such as these to the API, developers at Twitter make smaller refinements more frequently. For example, Axel Bruns and Yuxian Eugene Liang noted that an earlier version of the Stream API was not capable of capturing retweets if those retweets had been issued using Twitter's own retweet button found on its Web interface.[26] During this time, the Stream API only captured manual retweets, when the user had deliberately typed the requisite "RT" to indicate a retweet. Minor adjustments such as these can dramatically impact the yield of a data collection project; in the course of preservation work it is critical to retain enough information so that future users can be cognizant of these sorts of API behaviors. Finally, we can also acknowledge that as a commercial company and project, there is no guarantee that Twitter will always make its data available at all, or in the manner currently in use.[27] Given the potential of such an outcome, it is all the more pressing to document accurately the tools used to gather any given set of tweets.

Archiving a Tweet Collection to the OAIS Framework (ISO 14721:2012)

What is best practice for a set of tweets collected through the Twitter API? To determine this, we reference the ISO standard 14271:2012, also known as the Open Archive Information System (OAIS) Reference Model.[28] This is a broad framework that details the operations and terminology of an archives designed for long-term preservation of objects and is especially directed toward the preservation of digital objects. It should be emphasized that ISO 14271:2012 is neither an existing functional system nor an itemized checklist for such a system. Rather, it is a conceptual model for communities involved in long-term digital object preservation. As such, the recommendations here are not those of the standard itself, but are derived from the framework described there. Following this, we reflect on best practice for information gathering, but we do not recommend or investigate specific data structures (such as XML) or element sets (such as PREMIS or METS) for this information.

Here we will concern ourselves with the OAIS concept of the Archival Information Package (AIP). The AIP maintained by an OAIS system is regarded as the canonical version of the object, outside of any particular instance delivered to end users. The AIP consists of both the Content Information (in our case, the collected tweets) and the Preservation Description Information (PDI). Specifically, we will examine the five metadata concerns of the PDI: provenance, reference, fixity, context, and access rights information.

Provenance information

Provenance for a collection of tweets contains a twofold meaning. First, tweets come from Twitter users. This is well understood and conforms to the model of an author or creator. Second, collected tweets are derived from some specific methodological process, which must be mediated by a technological process as well. In our case, the technological process is the Twitter API, and the methodological process constitutes the chosen API commands, as well as the chosen arguments, such as keywords, hashtags, selected users, and the dates of execution.

The authorial provenance of tweets is problematized both by the vast number of tweets and authors to consider, as well as by the anonymity the platform affords. While Twitter provides a limited form of identity authentication for users who meet the criteria, it is not for general use by the Twitter community and is only applied selectively by the company for certain public figures. For researchers and archivists, then, it is difficult if not impossible to ascertain if any given Twitter account is used by single or multiple users, or to verify the identity of a user.[29]

Here we examine best practice for describing the provenance of the tweets as a dataset. As noted before, the specifics of the methodology behind collecting tweets are critical. Since a collection of tweets in any large quantity constitutes a dataset that could be described as big data, it is vital to guard against the misuses of such data and properly contextualize the dataset. Moreover, it is the general responsibility of any researcher to document methodology, and, in the case of Web and Twitter archives, this entails documenting the boundaries defined and the tools used to gather data within those defined boundaries.

In the ideal case, a record is produced of the API commands used with a timestamp of their execution. Any accompanying output of an API call should be recorded as well; this may include a count of the captured tweets and a count of filtered tweets that were not captured but match the search criteria. The resulting harvest for each command is then collected into a single grouping associated with that command. This may be manifested in a database table, spreadsheet, or other grouping strategy, such as a folder consisting of each returned tweet as a JavaScript Object Notation (JSON) file. Researchers will see fit to combine these subsets into a single set for analysis, but the subsets tied to their corresponding API executions should be retained. This practice conforms to the principle of provenance,[30] which asks that records of different origins (in this case, deriving from different API functions and times) remain unmixed to retain original order and context.

For example, the now defunct TwapperKeeper.com platform used a combination of both the Search and Stream APIs to fulfill a user's single search term request. This provided the user with a more inclusive yield to his or her query. The open source yourTwapperKeeper performs the same function but also adds the *archivesource* datapoint, which indicates for each tweet whether it was gathered through the Stream or Search API.[31] Although the program does not to our knowledge associate tweets with specific API executions, this is nevertheless an important feature. We recommend that similar care be taken with custom-made scripts and other capture methods. We believe this documentation provides excellent evidence of the limits of the API and of the corresponding dataset and can hedge against uncritical use of the data.

By default Twitter returns tweets as JSON objects. Researchers may need to transform this data structure to another for import into a database or other program. Transformations like these should be documented as well. Along with the potential for some data loss during such a process, such as the removal of metadata fields not of immediate concern to the researchers, the move to a new data structure, for example from JSON to a tabular data format such as CSV (comma-separated values), can remove the internal organization and hierarchy of the previous format. Although such information can be recoverable, we believe it best practice to retain the original returned format. In addition

to these concerns, more traditional provenance information such as chain of custody should also be recorded. Collectively, the provenance information will support the future authenticity of the dataset as well as the comparability of the findings.[32]

Reference and fixity information

Reference information functions as the identifier for the content. The specific reference values for the collection of tweets will depend on the particular system in use that provides unambiguous identifiers for archives objects. For a Twitter set, the primary decision is whether an identifier value will correspond to the dataset as a whole (i.e., the entire set of tweets gathered by the researchers for their project), to the subsets of tweets grouped by their API executions, to the individual tweets themselves, or to any combination thereof.

We anticipate the most useful and efficient approach to be mapping the identifier to the entirety of the dataset. Although associating an identifier for each API subgrouping of tweets would be valuable, it is less likely that those subsets will be frequently referenced, and relating the subsets systematically to each other introduces some complexity into any archival storage system. Associating an identifier with each tweet is also possible, but is contingent on retaining each tweet as a separate object, preferably as the JSON object originally received. Individual lines in a CSV file or entries in a database cannot be uniquely and permanently identified.

Fixity information "documents the mechanisms that ensure that the Content Information object has not been altered in an undocumented manner."[33] Its use will likely correspond with the chosen application of the reference information identifier. It should be noted, however, that there is more opportunity for granularity in running and recording integrity checks on components of the Twitter dataset, such as checks on individual JSON objects or subgroupings.

Context information

In the OAIS model, context information includes broader provenance concerns such as why the dataset or object was created and how it relates to other archives objects. In the case of a Twitter dataset, this space should include project information and background, such as the core research questions asked by the researchers. Most important for the context and future use of the dataset, it should include the reasoning behind the selection of keyword terms, hashtags, and other search parameters. Context information can also contain the researchers' "disciplinary standards, and cultural-contextual factors, as dimensions to consider in ethical decision-making."[34] This information will further

contextualize not just the circumstances of capture, it will also illuminate some of the culture and practice of the researchers' fields.

Access rights information

A number of authors and researchers have noted that the collection of social media data by researchers presents ethical concerns for the use and access of the data.[35] We identify two core issues regarding access. The first centers on the anonymity and safety of social media users when their postings are collected as data for analysis and sharing. The second issue centers on the ethics of using such data for analysis and research when no consent was explicitly given for such use.

We can acknowledge that in Twitter's present case there are no legal inhibitions for collection of Twitter data as users agree to such use through Twitter's "Terms of Service."[36] It can therefore be argued that the burden of anonymity and safety lies with the user. However, the collection and subsequent disbursement of Twitter data likely constitutes a unique and relatively new mode of use for which tweeps arguably do not provide informed consent as they traditionally would in other research contexts.

A discussion of the ethics of research in this context is beyond the scope of this article. We find it sufficient to note here that since the collection of informed consent is beyond practical reach in most cases of Twitter data collection, it is critical for researchers to consider the responsible and ethical use of this data and to document such considerations with the archived object. The Society of American Archivists *Code of Ethics*[37] states that archivists will protect the privacy rights of people or groups who are the subject of accessioned records. It therefore behooves archivists to consider the privacy concerns within any given tweet collection and to adopt an appropriate access model.

While access policies for a dataset will depend upon the institution(s) that host the data, the access rights information heading can indicate researchers' access concerns to future users of the collected tweets. These concerns may include a stipulation that the dataset not be mashed with other data points to maintain the anonymity of the users.

Conclusion: The Potential Uses of a Subject-Centered Collection

When the Library of Congress announced that it would provide long-term preservation for Twitter's entire archives of public tweets, Laura Campbell at the Library of Congress's Office of Strategic Initiatives wrote, "for historians, Twitter provides direct witness accounts of events in real time. It also serves as a virtual timeline of communications about events, people and places. This

provides an enormous amount of raw unmediated primary source material for historical research."[38] Twitter affords a unique opportunity to collect records of unfolding events unfiltered by mass media.

The implications for amassing and using subject-centered collections of tweets are enormous. Collections of tweets have been aggregated to determine how Twitter is used to disseminate information during and immediately after natural disasters.[39] A group of researchers performed content analysis on six thousand tweets from members of the United States Congress to determine their use behavior.[40] Another research team examined how Twitter is used to navigate imagined audiences among Twitter users who have become "micro-celebrities."[41]

If we are ever to determine how social media are used, we must have access to their content. Historians who examine the Arab Spring twenty years from now will be at a loss if they do not have access to the born-digital, primary source materials documented via Twitter. Collecting these materials is important because we have the unprecedented ability to collect not only records of human activity but the actual activity itself. While initial collection is vital, preservation of these datasets is equally so for long-term use. We have illustrated strategies for both of these activities. Although we believe they are sound, case studies using both the collection and preservation practices described in this article are needed to evaluate their feasibility and usability. To illustrate how Twitter was used to facilitate action on the ground, we provide examples in the appendix of protest tweets taken from a collection aggregated during the eighteen-day revolution in Egypt.

Appendix—Examples of Tweets from Tahrir

In this appendix, we examine a sample of tweets collected during the eighteen-day revolution in Egypt. These tweets were coded during a multi-institution study conducted by Timothy Arnold, one of the coauthors of this paper, and Matthew Rafalow and Amber Tierney from the Department of Sociology at the University of California, Irvine. The research team performed content analysis on a sample of tweets selected using a rigorous sampling strategy from a cache of over four million tweets written between January 25 and February 18, 2011, in Egypt and the United States. The study sought to explore how Twitter enables or constrains opportunities to develop collective identity and organize social movement activity locally, nationally, and transnationally. These tweets, their codes, and the definitions of the codes assigned during the aforementioned study are provided so that readers can understand more about the kind of data that could be preserved in a subject-centered collection as suggested in this article. Tweets are organized by code, which are in bold.

Collective Action: *Indicates the presence of some form of resistance (demonstration, protest, rally); noninstitutional.*

- RT @ AhmadFahmy: Millions protests planned for tomorrow isa. Possibly we'll protest in front of the State TV building "Maspero" #egypt
- heading towards downtown. Demonstrations at Egyptian Television Building, People's Assembly, Tahrir Square & other locations. #Jan25 #Egypt
- RT @aisaad: Call for March of Millions on Sunday, Tuesday, Friday! #Egypt #Jan25
- RT @Sarahngb: Translation RT @altahawi: people join us at Tahrir sq. Our numbers are sharply decreasing and we're exhausted #egypt
- PLEASE RT THIS ON YOUR TWITTER: #MubarakMustResign RT http://tinyurl.com/DigtialRally #EGYPT #DigitalRally #JAN25
- Please support https://secure.avaaz.org/en/egypt_blackout/ to help get #Egypt back online. If you can't donate please RT. Thanks
- Help fund more SAT uplinks. Donate $$ to help Tor get Non-state controlled #internet access to #Egypt! Please donate ---> http://ow.ly/3NJ6z

Repression/Censorship: *Indicates either that a particular statement is being censored or the act of censorship generally speaking.*

- RT @liamstack: i was just very briefly detained by army at egyptian museum. they let me go but mokhabarat [secret police] stole my camera. #Egypt #Jan25

- RT @ioerror: The internet service for TE Data in Egypt is being rate limited to 16KB/s - another form of censorship. #egypt #jan25
- RT @hamish6PM: Hotel security just entered our room and told us we are not allowed to have cameras on balcony #Egypt #6pm
- here is a chronological of communication shutdown that happened in #Egypt starting #Jan25 demonstrations http://bit.ly/hhwIR

Report of Corruption: *Indicates a report of corruption shared.*

- French PM admits that Mubarak paid for family holiday on Nile: http://t.co/PrbOzcQ (France now colony of Egypt & Tunesia?)
- RT @marmite_news: Video Confession #Jan25 Man paid £5000 by Ministry to wreak havoc in Cairo protests http://f24.my/dK0EOh #Egypt
- RT @3effat: In Tahrir square i SAW captured thugs admitting they were paid LE100 to protest & others with ID's of police officers #egypt
- RT @Port_Sa3eedy Holy crap, huge catapult made by thugs http://yfrog.com/5t990z #egypt #jan25 #tahrir
- Police ID from one of the goons #mubarak hired as police. "To protect and serve"? Not in HIS #Egypt http://bit.ly/fXYCHp
- # Video of how Egyptian police deal with demonstrators http://youtu.be/JbKUFEXxvhY #Egypt #Jan25 #Mubarak #Tahrir

Complicity: *Indicates or implies complicity of individual or state actors in the suppression of the uprisings.*

- Officer corps in Egypt has been trained in the US for the last 30 years
- The arguments the USA is using to control Egypt are the same as those by the Brits in their vain efforts to maintain their Empire.
- RT @guardiannews: UK refuses to suspend Egypt arms sales http://gu.com/p/2nx2c/tf
- US/UK Companies Help #Egypt Regime Shutdown Telecommunications & Identify Dissident Voices. DN!: http://ow.ly/3O9gO

Solidarity/Collective Identity: *Indicates a moment when culture is used in a way to indicate or encourage solidarity and reduce social distance between particular groups of people.*

- RT @nohaelshoky: Call your Member of parliament about Egypt.Why and how with UK,US & Canada links:http://www.twitpic.com/3v6r89 #Egypt
- RT @jocelyncarlisle: @monaeltahawy Pls RT: #NYC joins million man march for #Egypt Tue, 4-6 PM, Egyptian Consulate, 58th & 2nd. #Mubarak

- RT @demotix: International Egyptian Solidarity demonstration - Amsterdam http://t.co/NLqwA99 #solidarity #Egypt #Amsterdam
- RT @DominicKavakeb: On my way now to US embassy in London for #solidarity demo with #Egypt revolution #jan25 #tahrir

ICT-Proxy Access: *Indicates the user is providing access to phone, Internet, etc. through a proxy.*

- RT @wikileaks_pp: DNS ->; 8.8.8.8 | Twitter->; "128.242.240.52" Facebook->; "69.63.189.34" Google->; "72.14.204.99" #Egypt
- RT @embee: Helio: Mostafa 0103778585. Maadi: Marwa 0100057579. Mohd: Aya 0123337815. Zamalek: Dina 0123337815 Please circulate #Egypt

Notes

[1] Clay Shirky, "The Political Power of Social Media," *Foreign Affairs,* January/February 2011, http://www.foreignaffairs.com/articles/67038/clay-shirky/the-political-power-of-social-media; Ethan Zuckerman, "The First Twitter Revolution?," *Foreign Policy,* January 14, 2011, http://www.foreignpolicy.com/articles/2011/01/14/the_first_twitter_revolution.

[2] Ella Chou, "The Twitter Revolution Debate Is Dead," *The Atlantic,* February 14, 2011, http://www.theatlantic.com/technology/archive/2011/02/the-twitter-revolution-debate-is-dead/71185; Malcom Gladwell, "Small Change: Why the Revolution Will Not Be Tweeted, *The New Yorker,* October 4, 2010, http://www.newyorker.com/reporting/2010/10/04/101004fa_fact_gladwell; Anne Nelson, "The Limits of the 'Twitter Revolution,'" *The Guardian,* February 24, 2011, http://www.guardian.co.uk/commentisfree/cifamerica/2011/feb/24/digital-media-egypt; Dave Pell, "Egypt, Twitter and the Straw Man Revolution," *The Huffington Post,* January 30, 2011, http://www.huffingtonpost.com/dave-pell/egypt-twitter-and-the-str_b_815906.html.

[3] Luke Allnut, "Tunisia: Can We Please Stop Talking about 'Twitter Revolutions'?," *Radio Free Europe,* January 15, 2011, http://www.rferl.org/content/tunisia_can_we_please_stop_talking_about_twitter_revolutions/2277052.html.

[4] Jon B. Alterman, "The Revolution Will Not Be Tweeted," *The Washington Quarterly* 34, no. 4 (2011): 103–16.

[5] Laila Shereen Sakr, "A Digital Humanities Approach: Text, the Internet, and the Egyptian Uprising," *Middle East Critique* 22, no. 3 (2013), http://www.tandfonline.com/doi/abs/10.1080/19436149.2013.822241.

[6] Christian Kelleher, T-Kay Sangwand, Kevin Wood, and Yves Kamuronsi, "The Human Rights Documentation Initiative at the University of Texas Libraries," *New Review of Information Networking* 15, no. 2 (October 30, 2010): 94–109, http://www.tandfonline.com/doi/abs/10.1080/13614576.2010.528342.

[7] Kaplan Andreas and Haenlein Michael, "The Early Bird Catches the News: Nine Things You Should Know about Micro-blogging," *Business Horizons* 54, no. 2 (2011).

[8] Lee Humphreys, "Historicizing Microblogging," *Proceedings of CHI* (2010).

[9] Laura E. Campbell and Beth Dulabahn, "Digital Preservation: The Twitter Archives and NDIIPP," *Proceedings of iPRES 2010: Proceedings of the International Conference on Preservation of Digital Objects,* December 2010, http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/campbell-27.pdf.

[10] Richard Rogers, "Debanalizing Twitter: The Transformation of an Object of Study," *Proceedings of the 5th Annual ACM Web Science Conference*, 2013.

[11] Chris Messina, "Groups for Twitter; or A Proposal for Twitter Tag Channels," *FactoryCity*, August 25, 2007, http://factoryjoe.com/blog/2007/08/25/groups-for-twitter-or-a-proposal-for-twitter-tag-channels.

[12] Steven Vaughan-Nichols, "How Twitter Tweets Your Tweets with Open Source," *ZDNet*, August 30, 2012, http://www.zdnet.com/how-twitter-tweets-your-tweets-with-open-source-7000003526.

[13] Cory Doctorow, "Cyberwar Guide for the Iran Elections," *Boing Boing*, June 16, 2009, http://boingboing.net/2009/06/16/cyberwar-guide-for-i.html.

[14] Maureen Pennock, "Web-Archiving," *Digital Preservation Coalition Technology Watch Report,* March 2013, http://www.dpconline.org/advice/technology-watch-reports.

[15] "Frequently Asked Questions," https://archive.org/details/301works-faq.

[16] Stine Lomborg, "Researching Communicative Practice: Web Archiving in Qualitative Social Media Research," *Journal of Technology in Human Services* 30, nos. 3–4 (July 2012): 219–21, http://www.tandfonline.com/doi/abs/10.1080/15228835.2012.744719.

[17] Niels Brügger and Niels Ole Finnemann, "The Web and Digital Humanities: Theoretical and Methodological Concerns," *Journal of Broadcasting and Electronic Media* 57, no. 1 (2013): 66–80, http://www.tandfonline.com/doi/abs/10.1080/08838151.2012.761699.

[18] Lomborg, "Researching Communicative Practice, 219–31.

[19] Danah Boyd and Kate Crawford, "Six Provocations for Big Data" (paper presented at "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society," September 21, 2011), 1–17, http://ssrn.com/abstract=1926431.

[20] Brügger and Finnemann, "The Web and Digital Humanities."

[21] Brügger and Finnemann, "The Web and Digital Humanities."

[22] Fabian Neuhaus and Timothy Webmoor, "Agile Ethics for Massified Research and Visualization," *Information, Communication and Society* 15, no. 1 (2012): 43–65, http://dx.doi.org/10.1080/1369118X.2011.616519.

[23] Rogers, "Debanalizing Twitter."

[24] Biz Stone, "Finding a Perfect Match," *Twitter Blog*, July 13, 2008, https://blog.twitter.com/2008/finding-perfect-match.

[25] Mike Melanson, "Twitter Kills the API Whitelist: What It Means for Developers and Innovation," *ReadWrite,* February 11, 2011, http://readwrite.com/2011/02/11/twitter_kills_the_api_whitelist_what_it_means_for.

[26] Axel Bruns and Yuxian Eugene Liang, "Tools and Methods for Capturing Twitter Data during Natural Disasters," *First Monday* 17, no. 4 (2012), http://firstmonday.org/ojs/index.php/fm/article/view/3937/3193.

[27] Adam Edwards, William Housley, Matthew Williams, Luke Sloan, and Malcolm Williams, "Digital Social Research, Social Media and the Sociological Imagination: Surrogacy, Augmentation and Re-Orientation," *International Journal of Social Research Methodology* 16, no. 3 (2013): 245–60, http://www.tandfonline.com/doi/abs/10.1080/13645579.2013.774185.

[28] *ISO 14721:2012: Space Data and Information Transfer Systems—Open Archival Information System (OAIS)—Reference Model* (International Organization for Standardization, 2012).

[29] Heather Small, Kristine Kasianovitz, Ronald Blanford, and Ina Celaya, "What Your Tweets Tell Us about You: Identity, Ownership and Privacy of Twitter Data," *International Journal of Digital Curation* 7, no. 1 (2012): 174–97, http://ijdc.net/index.php/ijdc/article/view/214.

[30] Anne J. Gilliland-Swetland, "Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment," *Council on Library and Information Resources,* February 2000.

[31] Axel Bruns and Stefan Stieglitz, "Towards More Systematic Twitter Analysis: Metrics for Tweeting Activities," *International Journal of Social Research Methodology* 16, no. 2 (2013): 91–108, http://dx.doi.org/10.1080/13645579.2012.756095.

[32] Small et al., "What Your Tweets Tell Us about You."

[33] ISO 14721:2012.

[34] Stine Lomborg, "Personal Internet Archives and Ethics," *Research Ethics* 9, no. 1 (2012): 20–31, http://rea.sagepub.com/lookup/doi/10.1177/1747016112459450.

[35] Boyd and Crawford, "Six Provocations for Big Data"; Edwards et al., "Digital Social Research"; Fiona Gill and Catriona Elder, "Data and Archives: The Internet as Site and Subject," *International Journal of Social Research Methodology* 15, no. 4 (2012): 271–79, http://www.tandfonline.com/doi/abs/10.1080/13645579.2012.687595; Lomborg, "Personal Internet Archives and Ethics"; Neuhaus and Webmoor, "Agile Ethics for Massified Research and Visualization"; Steven Ovadia, "The Role of Big Data in the Social Sciences," *Behavioral and Social Sciences Librarian* 32, no. 2 (2013): 130–34, http://www.tandfonline.com/doi/abs/10.1080/01639269.2013.787274.; Small et al., "What Your Tweets Tell Us about You."

[36] Twitter, "Terms of Service," June 25, 2012, https://twitter.com/tos.

[37] Society of American Archivists, *Code of Ethics for Archivists*, 2011, http://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics.

[38] Campbell and Dulabahn, "Digital Preservation."

[39] Bruns and Liang, "Tools and Methods for Capturing Twitter Data during Natural Disasters."

[40] Jennifer Golbeck, Justin M. Grimes, and Anthony Rogers, "Twitter Use by the U.S. Congress," *Journal of the American Society for Information Science and Technology* (May 3, 2010), http://doi.wiley.com/10.1002/asi.21344.

[41] Alice E. Marwick and Danah Boyd, "I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience," *New Media and Society*, no. 1 (2010): 114–33, http://nms.sagepub.com/content/13/1/114.

ABOUT THE AUTHORS

Timothy Arnold is information manager at ENVIRON International Corporation. He has a dual master's degree from the School of Information and the Center for Middle Eastern Studies at the University of Texas at Austin. He has worked in knowledge management at a number of organizations including the United Nations. He is a graduate of Yale College where he received a bachelor's degree in history. He lives in Maine.

Walker Sampson serves as digital archivist at the University of Colorado Boulder. His research interests include digital object preservation and curation, with an emphasis on creative and aesthetic digital objects. He received his MS in information science at the University of Texas at Austin, where he participated in the IMLS funded "Preserving Games" research project. He previously served as electronic records analyst at the Mississippi Department of Archives and History.