CrossMark

API-based social media collecting as a form of web archiving

Justin Littman[1] · **Daniel Chudnov**[2] · **Daniel Kerchner**[1] ·
Christie Peterson[1] · **Yecheng Tan**[1] · **Rachel Trent**[1] · **Rajat Vij**[1] ·
Laura Wrubel[1]

Abstract Social media is increasingly a topic of study across a range of disciplines. Despite this popularity, current practices and open source tools for social media collecting do not adequately support today's scholars or support building robust collections for future researchers. We are continuing to develop and improve Social Feed Manager (SFM), an open source application assisting scholars collecting data from Twitter's API for their research. Based on our experience with SFM to date and the viewpoints of archivists and researchers, we are reconsidering assumptions about API-based social media collecting and identifying requirements to guide the application's further development. We suggest that aligning social media collecting with web archiving practices and tools addresses many of the most pressing needs of current and future scholars conducting quality social media research. In this paper, we consider the basis for these new requirements, describe in depth an alignment between social media collecting and web archiving, outline a technical approach for effecting this alignment, and show how the technical approach has been implemented in SFM.

Justin Littman is the lead author. All other authors contributed significantly to this work and participated in the writing of the paper. The authors are listed alphabetically. Laura Wrubel is the current principal investigator on the grant supporting this work; Daniel Chudnov held this role previously.

✉ Justin Littman
justinlittman@gwu.edu

1 GW Libraries, The George Washington University, 2130 H Street NW, Washington, DC 20052, USA

2 District Data Labs, 907 E Street SE, Washington, DC 20003, USA

1 Introduction

Social media is increasingly a topic of study across a range of disciplines. The content, communication patterns, technology, communities, and the roles and impact of social media have all emerged as areas of interest to researchers in fields from computer science and medicine to business, economics, and the humanities. Conferences such as the International Conference on Web and Social Media, Web Science, and Association of Internet Researchers provide substantial programs focused on web and social media research, methods, and data collection. Likewise, granting agencies such as the National Science Foundation (NSF) are supporting research both about and using social media.[1]

Social Feed Manager (SFM) is an application originally intended to support the needs of scholars collecting Twitter data for their research [1–3]. SFM was developed by a team at George Washington University (GW) Libraries[2] and has successfully supported scholars at GW and a number of other institutions since 2012. SFM uses the Twitter API to collect tweets from an identified user or to filter the stream of all tweets based on user, keywords, or geolocation.

Since its creation in 2012 to automate data collection for researchers, SFM has grown to become essential to the success of a wide variety of research projects at GW. GW Libraries has used SFM to collect tweets on behalf of faculty, graduate student researchers, archivists, librarians, undergraduate students, and non-faculty researchers. Scholarly

research has been conducted on tweets collected by SFM in a diverse array of disciplines including political science, media studies, business, women's studies, counterterrorism, and epidemiology, among others. For example, SFM has been used to

- Collect tweets to study the role of the gender of candidates for U.S. Congress in public responses to the candidate [4].
- Collect tweets from ISIS-affiliated individuals to analyze how ISIS uses Twitter to reach Western and non-Western audiences.
- Preserve the historical Twitter presence of the Corcoran School of Art and Design at the time of its merger with GWU.
- Provide a Twitter sample stream dataset (estimated 0.5% sample of all tweets) from which tweets with the #YesAllWomen hashtag were extracted to study Twitter's use in social activism.
- Enable students in a quantitative methods of political science course to easily gather tweets from members of Congress for analysis.
- Collect GW-related tweets, recognizing that social media content is now an indispensable component of a complete historical portrait of the contemporary GW community.
- Collect Sina Weibo content pertaining to China's anticorruption campaign for current and future China Studies scholars.
- Proactively collect 2016 presidential candidates' tweets, anticipating potential research value.

We anticipate that the need for tools to easily collect social media content in support of scholarly research and archives will likely increase commensurate with the expanding role of social media, whether on Twitter or other platforms, in societal communication and discourse.

Informed by this initial successful experience, the team is undertaking a complete rewrite of SFM supported by a grant from the National Historical Publications and Records Commission (NHPRC) [5,6]. This rewrite is additionally motivated by the continued development of social media research practices and the decision to expand the scope of SFM. Together, these have led us to reconsider our approach to social media collecting. In the new SFM, we are attempting to more closely align social media collecting with web archiving. As will be explained, not only has the web archiving community devoted attention to social media collecting, but it also has a rich set of practices and tools that are relevant to a technical approach to social media harvesting.

The goal of this paper was to make the case for more closely aligning API-based social media collecting with web archiving, using SFM to illustrate this alignment. To accomplish this, we will: (1) situate web archiving in relationship to social media collecting; (2) describe the relevant work in web archiving as it relates to social media collecting; (3) present some salient work in social media research related to research practices; (4) describe how SFM's goals have been expanded and how our thinking and assumptions about social media collecting have been reconsidered; (5) how this motivates requirements for a social media collecting application like SFM; (6) what it means to align social media collecting and web archiving and how this translates into a technical approach that satisfies the new requirements; and (7) how this technical approach has been implemented in SFM.

2 Situating web archiving and API-based social media collecting

The International Internet Preservation Consortium defines "web archiving" as "the process of gathering up (harvesting) data that have been published on the World Wide Web, storing it, ensuring the data are preserved in an archive, and making the collected data available for future research" [7]. Thus, web archiving is a broad concept in several aspects:

- Content: viz., anything on the Web (including social media)
- Digital content lifecycle stages: selection, harvesting, preservation, and access [8]
- Approaches: e.g., crawling vs. recording for harvesting:
- Technologies: e.g., Heritrix [9], Webrecorder.io [10], Warcprox [11]

One particular type of web archiving we will refer to is "traditional web archiving." By traditional web archiving, we mean retrieving and storing the pages of websites via HTTP, usually accomplished through the use of web crawlers. These web pages and their associated images, style sheets, JavaScript, etc., are stored in WARC files [12] and played back using wayback software.[3] Though traditional web archiving involves a number of different formats, the core format is HTML, which is intended to be human-usable when rendered by a web browser.

Traditional web archiving will be discussed in relationship to "social media collecting." By social media collecting, we mean retrieving and storing social media content from the web Application Programming Interfaces (APIs) of social media platforms. These APIs return data in a structured text format, typically JSON or XML, that is intended to be

---

[3] We refer here to "wayback software", a generic term for software that plays back WARC files, as distinguished from "The Wayback Machine", an instance and implementation of wayback software hosted by the Internet Archive. Two examples of wayback software are the International Internet Preservation Consortium's OpenWayback [13] and Ilya Kreymer's pywb [14].

Table 1 Distinctions between social media collecting and traditional web archiving

	Primary record unit	Primary response format	Source
Social media collecting	Post (with metadata including links to associated resources)	JSON or XML	API
Traditional web archiving	Web page (with embedded links to associated resources)	HTML	Website

machine-readable. As with HTML pages, any non-text elements of social media posts are typically referenced in the JSON/XML by URIs from which they can be linked to or retrieved and rendered with the content.

Harvesting from the API is one of several possible approaches to collecting social media data. Other approaches include the following: purchasing from data resellers; using a commercial, third-party service; using a platform self-archiving service; and harvesting with web crawlers [15]. Compared to crawling social media web interfaces, some of the advantages of harvesting from the API include the following:

- The data are structured (typically JSON or XML). This is essential for research that involves applying computational techniques.
- To avoid disrupting third-party applications, social media platforms tend to keep their APIs stable.
- Some social media platforms provide more metadata via their APIs than they do via their websites.
- Data can generally be collected more efficiently.

Harvesting from social media APIs has a number of disadvantages as well that must be considered:

- Not all social media platforms have complete, public APIs. For example, one of our discoveries from working with the Sina Weibo API is that the data that are available without approval from Sina Weibo are extremely limited.
- Data are not readily human-viewable.
- Each API is different.
- Currently, there is no standard accepted format for the archival storage of social media data, although our work described here explores WARC as a common container format for data from a variety of social media platforms.
- Some platforms, notably Twitter, place explicit requirements and limits on the use and sharing of data harvested using the API.
- Limitations placed by the platform on the amount of data that can be harvested via the API can make it difficult or impossible to capture older content.

As will be explained, we consider social media collecting as a form of web archiving. Table 1 summarizes the salient distinctions between social media collecting and traditional web archiving.

3 Background and related work

3.1 Web archiving

Archival institutions have been collecting, managing, and making accessible social media content since the advent of GeoCities, Blogger, and other early social media platforms of the 1990s, even before the term "social media" was in general use [16]. The Internet Archive's oldest GeoCities harvests date back to 1996[4] (one year after GeoCities' founding); its oldest captures of Blogger,[5] Friendster,[6] MySpace,[7] and Flickr[8] date back to each of these platform's founding (1999, 2002, 2003, and 2004, respectively).

As sites like Flickr, YouTube, Twitter, Facebook, and other social media platforms with public APIs grew on the Web in the following years, the members of the archival community began collecting these too, either through broader web archiving programs or, eventually, through archiving programs that specifically targeted social media—sometimes referred to as "social networking archives" or "web 2.0 archives." Early programs that explicitly targeted social media were often within governmental archives. In 2008, The National Archives and the Internet Memory Foundation began a pilot program to capture social media content in the UK Government Web Archive using API-based harvesters [17]. In 2009, the State Archives of North Carolina and State Library of North Carolina officially expanded the scope of the North Carolina State Government Web Archives to include state agency social media accounts, putting to use Archive-It's Heritrix web crawler to capture Twitter, Facebook, YouTube, Flickr, and (eventually) additional platforms [18–21]. In 2010, the Library of Congress signed an agreement with Twitter to archive the entire body of tweets ever made since 2006, inspiring a public discussion around pri-

---

4 https://web.archive.org/web/*/http://geocities.com.

5 https://web.archive.org/web/*/http://blogger.com.

6 https://web.archive.org/web/*/http://friendster.com.

7 Although the URL "http://myspace.com" was captured from 1996 forward, MySpace was founded and launched at that URL in 2003. https://web.archive.org/web/20031004101518/http://myspace.com/.

8 https://web.archive.org/web/*/https://www.flickr.com/.

vacy, copyright, and the value of the public historical record on Twitter [22–24].

The National Digital Stewardship Alliance conducted its first survey of the web archiving field in 2011, finding that 38% of respondents included or planned to include social media in their archives. Examples of existing collections reported by respondents included collections of human rights material, labor movements, leftist activism, and institutional records [25]. Similar social media collections from 2011 and 2012 included those documenting the Occupy movement [26–28], Hurricane Sandy [28,29], and campus controversy [30].

Although web archiving institutions were increasing efforts to capture, archive, and make available the rich historical record being created on social media, archivists were struggling with a lack of technical standards for storing social media and choosing amongst an array of tools that were unable to capture many of the most important aspects of social media: its discursive layers (replies, likes, comments, favorites, hashtags); underlying modular structure (datasets that can be queried and repurposed); and connection to the rest of the web (links to websites, and vice-versa). Existing collection methodologies ranged from capturing social media content through web crawlers, donation of record creators' own social media exports via a platform's self-archiving service, the use of third-party RSS readers and commercial API harvesters, and obtaining data directly from the platform organizations or data resellers [15,31].

The architects of the UK Government Social Media Archive described how existing web crawlers were inadequate to the complexity of capturing social media data, requiring their 2008 pilot program to develop new solutions to this complex problem [17]:

> [Archiving social media] requires specific methods and new models for implementing the complete preservation workflow: from the harvesting tools and methods to the access and presentation requirements. The traditional harvesting techniques based on parsing of the webpages and explicit link extraction will not reliably succeed in retrieving the complete content, especially when it is dynamic and provided at the discretion of the content sources and service providers.

In 2012, the State Archives of North Carolina and the State Library of North Carolina announced their adoption of ArchiveSocial, a commercial API-based capture service implemented in tandem with the Archive-It crawler captures of social media content [21,32]. Since 2012, the North Carolina program has been capturing state agency social media accounts as web-crawled WARCs (from Archive-It) and a subset of those accounts as JSON harvests (from ArchiveSo-

cial).[9] Around the same time, similar efforts began on the development of open source API-based solutions: Emory Libraries Digital Software Commons developed Twap [33] to facilitate capture of Occupy tweets, North Carolina State University Libraries developed Lentil [34] for capture of Instagram content, and GW Libraries began working to make SFM more useful to archivists.

Despite these advances, the field of social media harvesting tools is still largely unaligned with web archiving technology—in terms of capture, storage, format standards, access and reuse, and legal restrictions. The need for such an alignment is widely felt by web archivists, as are the pressing challenges (see for example [35,36]). The ethical, technical, and logistical challenges of social media archiving have been represented at every meeting of the Society of American Archivists since 2010. The National Digital Stewardship Alliance's 2013 survey of the state of web archiving in the United States found that the content type respondents were most concerned about being able to archive was social media.[10]

3.2 Scholarly social media research

While social media data have become a critical source for social scientists and researchers in a wide range of disciplines, social media data present particular challenges to scholarly researchers in fully documenting their methodology and supporting replicability and reproducibility of their results [38,39]. In a qualitative, ethnographic study of social media researchers, Weller and Kinder-Kurlanda explore current practices in social media research [38]. They point out that a number of the barriers to quality research exist due to a lack of documentation within the research process, especially data collection: "Being able to retrace all steps of collecting, processing and cleaning the data was seen as crucial for data quality and ensuring that the data really held what it promised". With greater understanding of the problem, current practice is evolving, yet "researchers admitted that they already had experienced difficulties in keeping track of their own activities and understanding what exactly they had done in order to collect or clean a specific dataset".

Weller and Kinder-Kurlanda also identify a set of issues, referred to as "data collection problems," that are specific to social media research. As they explain, "Many researchers faced problems with data quality, such as a lack of clarity

---

[9] ArchiveSocial requires social media account owners to login and give ArchiveSocial permission to their social media data. One of the authors of the paper worked with the adoption of ArchiveSocial at the State Archives of North Carolina.

[10] Noting also that, "The research, development, and technical experimentation necessary to advance the archiving tools on these fronts will not come from the majority of web archiving organizations with their fractional staff time commitments" [37].

with regard to the bias in the sample provided by the API, insufficient documentation of the data, and opaque collection processes". Sara Day Thompson and William Kilbride have pointed out that other commercial social media data providers such as Gnip and Datasift do not provide information about how they have selected data from the API datastream [34]. It bears noting that this lack of clarity applies to the decisions the social media company has made about what data to make available via the API and what may be excluded.

Incorrect inferences and interpretations of a platform's data are also a risk in social media research, as these depend on a clear understanding of its mechanics and usage, how these change over time, and likewise the broader ecology in which these actions occur [40]. Preservation and documentation of each platform's usage practices, its API responses' fields, and its web and mobile interfaces are complementary work to data collection and in need of attention to support research.

3.3 Non-academic social media research and archives

Social media data are also widely and actively used by the legal community, journalists, government agencies, open government advocates and seekers of government records, institutional records managers, and other "non-academic" researchers [41,42]. Similar to academic users, these researchers face challenges when using social media data collections that lack appropriate documentation of how the data were collected and managed over time. In a 2014 study of public relations professionals and the challenges they face preserving social media records, Hajtnik et al. argue that "organizations (and/or public relations practitioners) need to make sure that the public relations records they create for social media are preserved in a way that makes them accessible, usable and authentic for eventual later use and for future research." Pointing to established rules of archival practice and legal precedent, Hajtnika et al. find that it is crucial that institutions be able to demonstrate the authenticity of records through documentation of steps taken in collecting and managing social media data.

The concept of authenticity—that a record is what it purports to be—and its associated concept of integrity—that a record has not been altered or tampered with in a way that undermines its authenticity—are fundamental to archival practice and the evidential use of digital records in American courts [43,44].

3.4 Ethical considerations in social media research

At all stages of work with social media data there are ethical considerations, especially concerning privacy, consent, harm to human subjects, and data access and reuse. Even with social media that is considered "public," these factors apply. Understanding of these concerns and researcher practices are evolving and receiving increasing attention in the form of professional ethical frameworks and broader discussion [45–49]. How these aspects might be embedded effectively into data collection tools and archival workflows is an area of inquiry to which the SFM project team is attuned.

4 Reconsidering social media harvesting with SFM

4.1 Supporting building collections

SFM was originally developed to fill on-demand requests for Twitter datasets from on-campus scholarly researchers. The tool was inspired by conversations with a faculty member who had tasked her students with manually cutting and pasting tweets into a spreadsheet over the course of a semester, due to a lack of better, easily available tools. In our typical use case for the original SFM, a researcher—usually a faculty member or a student—approaches GW Libraries about collecting some tweets for a research project. The team member configures SFM to collect the data, and after some period of time, exports the tweets to Excel/CSV and delivers to the researcher. In other cases, the team member delivers JSON data and works with the researcher to extract the relevant fields. Depending on the research methodology, the researcher might load it into her own analytic tool such as Stata, SPSS, or R; process with some custom software; and/or code individual tweets according to conventions specific to their research and appropriate to their discipline. For this use case, we are serving today's social media researchers by constructing a dataset to meet their specific research question and requirements.

With the rewrite of SFM, we are adding a family of use cases: collection builders creating collections of social media data from which future researchers can extract datasets (It should be noted that we mean "researcher" in the broadest possible sense as anyone with an information need that involves social media data; that person may or may not be an academic). The shift is from collecting in response to a specific research question for a very specific use to speculatively collecting for many possible uses in research. It also introduces a distinction between the person doing the collecting (such as an archivist) and the person doing the research (the researcher). The activities of the collection builder and researcher are also likely to be separated by time; for historians, for example, that amount of time may be quite significant.

Of course, building collections for future researchers is the general approach of web archiving institutions such

as the Internet Archive[11] and the Library of Congress[12] and archives more generally. This shift in data collecting approach has made the work on SFM relevant to conversations within the archival community about how to best capture and make available social media materials. In fact, one of the key components of the NHPRC grant [5,6] is engaging archivists to inform the development of the new SFM and explore the intersection between social media harvesting and archival practice.

The timing of this shift to support building collections is fortuitous, as it coincides with an emerging discussion between the web archiving community and historians. With twenty plus years behind it, the earliest web archives are now of active interest for historical research [50]. Historians such as Ian Milligan and Peter Webster are providing valuable feedback on how web archives can better support historical

media platforms are likely to come and go.[13] The NHPRC grant calls for adding support for Flickr and Tumblr; a separate grant from the Council of East Asian Libraries[14] funds support for Sina Weibo.

4.3 Social media is more than text

Social media is often rich with links to other web resources, such as web pages, images, audio, or video [15]. These web resources are candidates for collecting as complements to the social media.[15]

The need for collecting related web resources is particularly evident for social media platforms such as Flickr and Tumblr, as they are not primarily textual. For Flickr, the primary unit of content is a photo. For Tumblr, the primary unit is a blog post, but that post can be text, image, audio, video, a link, or a chat. Here is an excerpt from a Tumblr video post:

```
{
        "id": 113915730303,
        "post_url": "http://justinlittman-
dev.tumblr.com/post/113915730303/test-post-5",
        "type": "video",
        "video": {
                "youtube": {
                        "video_id": "p3XgpvO2t5g",
                        "width": 540,
                        "height": 304
                }
        },
        "player": [{
                "width": 250,
                "embed_code": "<iframe width=\"250\" height=\"141\"
id=\"youtube_iframe\"
src=\"https://www.youtube.com/embed/p3XgpvO2t5g?feature=oembed&amp;enablejsap
i=1&amp;origin=https://safe.txmblr.com&amp;wmode=opaque\" frameborder=\"0\"
allowfullscreen></iframe>"
        }]
}
```

research [51–53]. What they say about web archives is likely to have implications for social media archives.

4.2 Moving beyond Twitter

Though Twitter provides one of the richest troves of social media data today, it is far from the only social media platform of value to scholars and, as history has shown, social

---

[11] https://archive.org/.

[12] http://www.loc.gov/webarchiving/.

[13] Many of us remember Friendster, MySpace and other extinct social platforms. Though certainly more popular, even Twitter itself seems to be experiencing a stall in the growth of its user base [54].

[14] The GW Libraries are collaborating with Johns Hopkins University and Georgetown University in this grant work, entitled "Blogging and Microblogging: Preserving Non-Official Voices in China's Anti-Corruption Campaign".

[15] Another aspect of tweets is the metadata that accompanies it when harvested from the API. This metadata contains social network information, in that they contain references to (and/or retweets of) other accounts. In addition, tweets contain complete user profile information, which often changes over time. This metadata has research potential, which is why we have also saved it.

Even social media that is primarily textual such as Twitter is ripe with links to other web resources. Here's an excerpt from a tweet:

```
"urls": [{
    "url": "https:\/\/t.co\/KV9npvzrUW",
    "expanded_url": "http:\/\/gwu-libraries.github.io\/sfm-ui\/posts\/2016-
07-11-pulse-processing",
    "display_url": "gwu-libraries.github.io\/sfm-ui\/posts\/2\u2026",
    "indices": [82, 105]
}]
```

The value of the related web resources is evident in the work performed with the Ferguson Twitter archive, e.g., Ed Summers's extract of URLs from the archive [55] or Ryan Baumann's download of Ferguson-related videos [56]. The value of these links is further reinforced by the case made by Milligan and Nick Ruest that constructing a web archive from the web resources linked to from tweets can complement web archives curated by archivists [57].

4.4 Quality research and archives require provenance metadata

As discussed above in Sect. 3.2, one common barrier to producing valid and replicable research with social media data that social media researchers have identified is a lack of documentation about the data collection process. And, as discussed in Sect. 3.3, non-academic researchers also require documentation about the data collection process to establish the trustworthiness of the data.

This type of research documentation is closely aligned with the archival field's concept of provenance, a concept foundational to the field and referring to information that traces the origin and chain of custody of archived information. In the context of digital curation, there is a particular emphasis on tracing provenance back to the origination of the data, including the particular settings and configurations used to collect and create data [58,59].

The W3C's PROV working group offers a definition of provenance that provides a common ground for the social media research community and the archival community [60]:

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

Through our work with Twitter data in particular, the SFM team has identified three distinct categories of provenance-related activities in the creation of a collection [61]:

- Creation: the creation of the social media post itself.

- Collection: what was collected, by what means, and on what schedules.

- Selection (or, in archival terminology, "appraisal"): the human process of deciding which social media to collect and which not to collect, that is, the decisions that guide the collection action.

Creation information comes directly from the social media platform and forms part of the data that are being gathered. It may include information about the author, such as screen name and date the account was created, as well as information about the post itself, such as the date, time, method, and location of posting.

In the SFM system, collection information is stored in the WARC files that are output by the collection process, as well as several related tables in the SFM database. The WARC files record the actual HTTP requests and responses by which the data are collected, while database fields contain information that would not normally be present in the WARC files, but which might be important to provenance, such as informational, warning, or error messages received during the collection process.

Finally, selection information is stored in SFM through an automatically generated change log supplemented by collection description and notes that can be added to most actions. So, for example, the change log might reflect that on a particular date, three Twitter handles were added to a collection and five were removed. Notes could be added to each of these actions to indicate, for example, that a particular user was removed from collection because she is no longer tweeting about the subject of the collection.

While the SFM team is convinced that provenance metadata is important to the support of quality research and a historical record that is authentic and reliable, the sheer quantity of potential provenance metadata can be overwhelming. We are endeavoring to engage researchers, archivists and others in a conversation to help us determine which fields to feature and privilege in SFM, and which may be hidden by default.

4.5 Putting harvesting tools in the hands of collection builders

Initially, SFM required researchers to ask system administrators to create the seeds and harvests on their behalf. Without the assistance of a system administrator, users were limited to viewing and downloading collected tweets. Based on practical experience and as we have reconsidered the roles around social media collecting, it has become clear that more of a "self-service" application will better meet user needs. This would allow users to initiate and tweak harvests, and request exports of various types themselves. This approach is similar to services such as Archive-It, webrecorder.io, and ArchiveSocial.

At the same time, we are aware that social media collecting often requires a deep understanding of the APIs and the tool doing the collecting. Thus, the self-service aspect must be balanced with continued support for users in their work with SFM.

5 New requirements for social media collecting with SFM

The focus of the original SFM was collecting and providing exports of datasets of tweets. Based on our experiences with SFM, the reconsideration of social media collecting just described, and listening to social media researchers, archivists, and potential future researchers, a number of additional requirements for social media collecting with SFM emerged:

1. SFM should be able to collect heterogeneous content from multiple social media platforms and from the Web.
2. SFM must record provenance metadata.
3. SFM must be usable by a variety of types of users without mediation.

Partly motivated by the principle that digital preservation necessitates access, and partly motivated by intuition that Web-like access will be necessary for the sort of research some future researcher will want to perform, we propose a fourth, speculative requirement:

4. SFM should be able to support access to collected content in a "Web-like way," including links to related web resources.

Distinct from exporting the content as datasets, Web-like access allows a user to interact with a rendered form of the content. This would support various forms of browsing and discovery, as well as navigating between the social media content and related web resources and embedded resources

(e.g., images, video, and audio). The specifics of this requirement are not yet clear at this stage in the project. For example, what, if any, of the functionality or look-and-feel of the original social media website should be replicated? What sort of discovery is useful or possible? Despite the speculative nature, we are confident that some sort of web-like access is a requirement for social media collecting.

While discussing requirements, it is worth being explicit about two requirements that are not in scope for SFM: first, SFM does not provide analytics for social media data. Based on our experience with researchers, the proper role for SFM is to provide exports of datasets so that they can be loaded into the researcher's own analytic tools. The minor exception to this is to provide some minimal analytics or discovery to support collection building (This is described further below). Second, SFM is not a preservation platform. Our goal is to create preservable data, but we assume that like other software, over the long-term, SFM will come and go. The provenance and export requirements entail being able to export social media data from SFM into an institution's preservation environment.

6 Technical approach

With the context of our reconsiderations of social media collecting and updated requirements described above, it is clear that API-based social media collecting is not a wholly distinct undertaking. Rather, it is a form of web archiving and bears some striking resemblances to traditional web archiving. We hypothesize that taking a technical approach that aligns social media collecting with traditional web archiving practices will allow these requirements to be satisfied.

The key to aligning social media harvesting and traditional web archiving is recognizing two similarities. First, irrespective of whether a crawler is requesting a web page from a website or a social media harvester is requesting social media data from an API, the entire transaction occurs over HTTP (Fig. 1).

Second, both web pages and social media posts contain links in the form of URIs to other web resources. In both cases, the linked web resources may be candidates for collecting as well.

Based on these similarities, the outline of a new technical approach can be proposed:

- Use WARCs to store harvested social media content and related web resources.
- Use a web proxy to create WARCs.
- Use existing web crawlers for harvesting related web resources.
- Play back social media content and related web resources using wayback software.
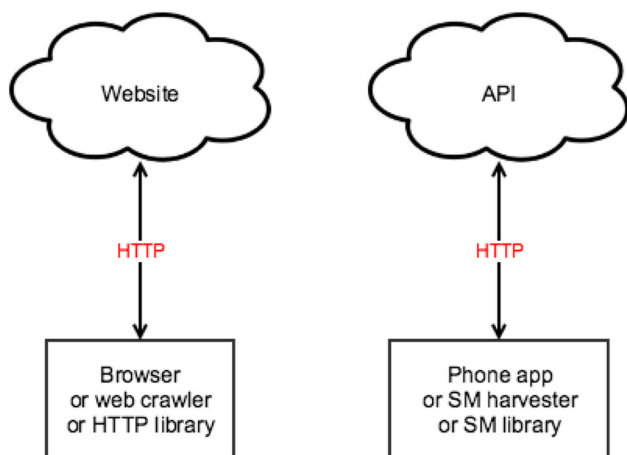
Fig. 1 Traditional web archiving and social media harvesting

This represents only an outline; other opportunities for leveraging web harvesting tools and practices remain to be considered.

6.1 WARC files as containers for social media data

WARC (or Web ARChive) files are intended to be "a container format that permits one file simply and safely to carry a very large number of constituent data objects" of "unrestricted type (including many binary types for audio, CAD, compressed files, etc.)" while only having "minimal knowledge of the nature of the objects" [12]. What distinguishes WARC from other container formats such as ZIP or TAR are features for storing not only the data objects, but also the complete messages exchanged as part of an HTTP transaction. As the WARC Specification states, WARC is designed for the following:

- to store both the payload content and control information from mainstream Internet application layer protocols, such as HTTP, DNS, and FTP;
- to store all control information from the harvesting protocol (e.g., request headers), not just response information.

WARC files are typically used by traditional web archiving for storing collected websites and all of the rich web resources that this entails.

WARC files provide a number of valuable features for social media harvesting. First, WARC provides a single storage format for social media content from all platforms and for all related Web resources. It avoids a proliferation of storage formats as new social media platforms are added or new types of related web resources (e.g., photos, video) are collected.

Second, WARCs help fulfill the requirement to record provenance metadata by recording the HTTP messages exchanged between the social media client and the API. This
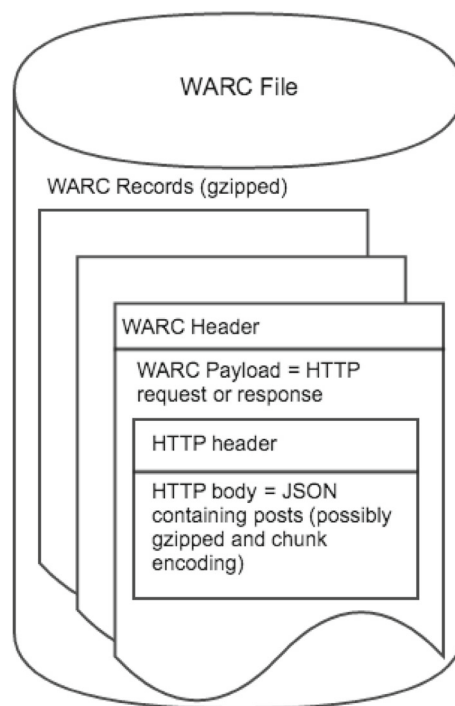

Fig. 2 Social media data stored within a WARC file

includes not only the social media data exactly as it was transmitted, but also other critical metadata such as the request that was made to the API (which includes which API method was called, what parameters were provided, which credentials were used), the date and time of the request and response, and the identity of the server that provided the response.

Compared to simpler storage formats for social media data such as line-oriented JSON files, storing social media data in WARC files introduces some additional complexity. Unlike a line-oriented JSON file, which may simply be gzipped (a type of compression), social media data in a WARC is nested in multiple layers of encoding. The actual social media data will be in the body of an HTTP message which may have gzip content encoding and chunk transfer encoding; the HTTP message will be in a WARC record which may be gzipped; and the WARC record will be in a WARC file. This is illustrated by Fig. 2.

Though as described below, this complexity is a factor in considering this approach, we have produced tools that abstract this complexity and provide access to the social media data stored in WARCs.

Another aspect of using WARC files for storing social media data is that it is largely "read-only". A by-product of recording the exact HTTP transaction is that the social media data cannot be modified, e.g., to delete specific social media posts (e.g., for privacy reasons) or enhance the social media posts with additional metadata (e.g., to unshorten URLs). Thus, this approach may not be a good fit for uses which
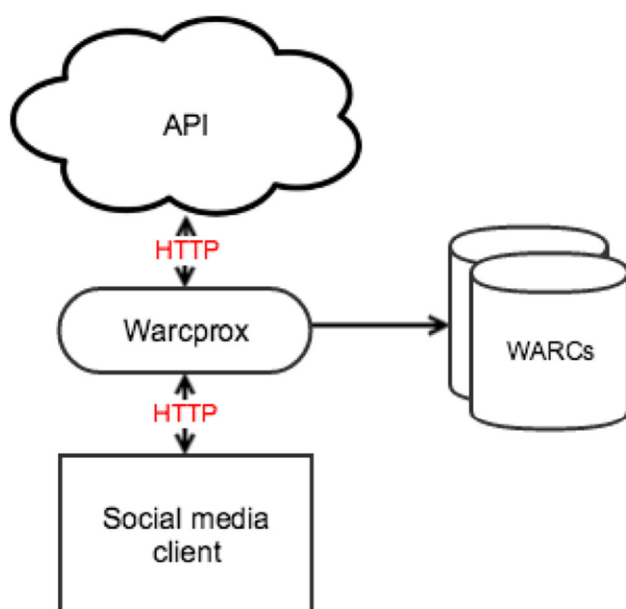
Fig. 3 Web proxy for creating WARCs

require curation at the level of individual posts. Alternatively, it may require that curation occur outside of the WARC files and/or upon or after export from the WARC files.

6.2 Use a web proxy to create WARCs

A web proxy is one of the main strategies for archiving websites. It is used by web archiving tools such as webrecorder.io and brozzler [10,62]. Though there are varying implementations, the basic approach is to interject a web proxy between an HTTP client and a web server. The web proxy relays HTTP messages between the two, taking advantage of this positioning as a "man-in-the-middle" to record the HTTP transaction to a WARC file. This is illustrated by Warcprox in Fig. 3.

Creating WARC files using a web proxy offers a number of advantages over other possible approaches [63]. First, because of its positioning between the HTTP client and the server, it allows recording the exact HTTP transaction without modification. Second, since the HTTP client does not need to be aware of its role in web harvesting, existing social media API clients, e.g., Twarc for Twitter, can be used. Thus, a social media harvester can be constructed from an existing tool for recording to WARC files and existing clients for interacting with social media APIs, minimizing the amount of original software development and allowing the leveraging of proven, widely used code.

6.3 Use existing web crawlers for harvesting related web resources

The third point in the outlined approach is to use existing web crawlers to harvest related web resources. The responsibil-

ity of a social media harvester then becomes retrieving data from the social media API and extracting the links to related web resources (Depending on how the social media platform structures its data, this can be easy or hard. For Twitter, links are already parsed from the tweet and delivered in the metadata; for Tumblr, some links must be extracted from HTML snippets). Rather than reinventing harvesters for web pages, video, and the like, that work can be delegated to existing web harvesting tools such as Heritrix and Wpull [64].

6.4 Play back social media content and related web resources using wayback machine

Recording the social media data and the related web resources as HTTP transactions in WARCs allows for the possibility of playback using wayback software. While wayback software can function in several modes [13], in the most common mode it returns an archived web resource to a requester based on the original URL and a target date. For web resources that contain links, the wayback software rewrites embedded URLs so that they reference the wayback instance ("archival URLs") instead of the original target. As a result, a user accessing an archived website from a web browser interacts with a rendered version of the web resource as it existed at the time of capture. To varying degrees, this will preserve the look-and-feel and functionality of the historical web resource.

Playing back social media data with wayback software provides a time- and URL-aware API to the social media content just as it does for web pages. The API allows requests to be made such as "give me the capture of gelmanlibrary's user timeline closest to January 1, 2016." As evidenced by the Memento "Time Travel for the Web" protocol (RFC7089), a time- and URL-aware API is a central means of accessing web archives [65]. Further, wayback software provides a mechanism for resolving links between social media content and related web resources.

It is important to recognize the difference between the playback of web pages and the playback of social media data. The source of the social media is the API, not the social media platform's website. It is in JSON format (or structured text format), not in HTML; when replayed, it will not have the look-and-feel or functionality of the social media website. By itself, this will not satisfy the "Web-like" access requirement, but provides a foundation upon which applications that render the social media data could be built.

7 Implementation

In June 2016, we released version 1.0 of SFM [66]. While not yet manifesting the entire technical approach just described, it implemented enough of this approach to provide evidence

Fig. 4 Creating and updating collections

of the viability of adopting web archiving practices and tools for social media collecting. The following provides an overview of the initial release of SFM, as well as some related experimental work.

SFM is implemented as a set of loosely coupled services written primarily in Python. The main service that a user interacts with is SFM UI, which is a Django web application that allows creating and curating collections, scheduling harvests, and requesting exports. There are separate services for harvesting and exporting each of the social media platforms. The services communicate via JSON messages exchanged using a messaging queue (RabbitMQ). Data and service state are persisted to a shared filesystem.

Services are containerized in one or more Docker images. Docker is a technology for packaging systems such as SFM to simplify deployment to multiple environments (e.g., Amazon web services) in multiple configurations (e.g., development or production). Our purpose in using Docker is to drive down the barriers for institutions deploying SFM.

7.1 Defining, describing, and organizing collections

Using the SFM UI, users can specify what to harvest by creating and updating collections.

Each collection has a harvest type. Collection harvest types differ based on the social media platform and the part of the API from which the social media is to be collected. For

Fig. 5 Adding a collection

example, a "Twitter search" collects tweets from Twitter's search API.

In version 1.0, the harvest types supported by SFM include Twitter search, Twitter filter stream, Twitter user timeline, Twitter sample stream, Flickr user, and Weibo timeline [67–70]. Tumblr blogs were added in a subsequent release (Figs. 4, 5, 6; Table 2).

SFM allows the user to create multiple collections of each type within a collection set. For example, the user might create a "Democratic candidate Twitter user timelines" col-

Fig. 6 Seeds

Table 2 Harvest types and seeds

Harvest type	Seed	How many?
Twitter search	Search query	1 or more
Twitter filter	Track/follow/locations	1 or more
Twitter user timeline	Twitter account name or ID	1 or more
Twitter sample	None	None
Flickr user	Flickr account name or ID	1 or more
Weibo timeline	None	None

lection and a "Republican candidate Twitter user timelines" collection. Collections are one way of organizing harvested content.

Each collection's harvest type has specific options, which may include the following:

- Schedule of how often to collect (e.g., daily, monthly). Harvests from Twitter's streaming API do not have a schedule—they are either on or off.
- Whether to perform web harvests of images, videos, or web pages embedded or linked from the posts.
- Whether to harvest incrementally. For example, each time a Twitter user timeline harvest runs, it can either collect only new items since the last harvest or it can recollect each entire timeline.

Some harvest types require seeds, which are the specific targets for collection.

The definition of a seed and the number of possible seeds varies by harvest type.

Note that some harvest types do not have any seeds. The Twitter sample harvest type collects a random sample of public tweets; there is no control over what is collected and hence no seeds. For the Weibo timeline harvest type, the weibos that are collected are determined by the friends of the user whose credentials are used; friends are set on the Sina Weibo website, not from within SFM.

Collections can be organized into collection sets. For example, the "Democratic candidate Twitter user timelines" collection and a "Republican candidate Twitter user timelines" collections may be placed in the "2016 Election" collection set. Collection sets are owned by a group, to which multiple users may belong. This allows collection responsibilities to be shared across a team.

In addition to defining and organizing collections, users can describe collections and collection sets by providing a name and a description. The description allows users to document aspects of their curatorial intent such as organization and selection criteria.

A log is kept in the database of each change to a collection set, collection, or seed. The log automatically includes the time of the change, the user that performed the change, the fields that were changed, and optionally may include a note from the user describing the change. Figure 7 shows a change made to the schedule of the "Democratic candidates" collection.

The description fields and the change logs are one aspect of the provenance metadata recorded by SFM.

7.2 Harvesting

Another function of SFM UI is to send harvest request messages to the messaging queue. Harvest request messages contain all of the information needed by a harvester to perform a harvest such as the harvest type, seeds, and credentials. Harvest request messages are either sent according to a schedule, or for streaming harvest types, when the user turns them on.

There are separate harvesters for each social media platform, but they all follow the same steps. The SFM architecture is open to the contribution of additional harvesters in the future.

Upon receiving a harvest request message, a harvester does the following:

1. Launches an instance of warcprox. Warcprox is a web proxy created by Internet Archive that records the HTTP

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

Fig. 7 Change log

Table 3 Social media clients

Platform	Client
Twitter	twarc [71]
Flickr	flickrapi [72]
Sina Weibo	Client written by SFM team
Tumblr	Client written by SFM team

transactions to WARC files [11]. Warcprox records the HTTP transaction between the social media client and the API.

2. Invokes a social media client to make API requests based on the harvest request. Table 3 shows the social media clients used by SFM.
3. Extracts links and update counts as responses are received from the social media client.

For example, if the harvest request type is a "Twitter user timeline," then the harvester will call Twarc's timeline() method, which will call https://api.twitter.com/1.1/statuses/user_timeline.json, which is the user timeline method in the Twitter REST API.[16]

Upon completion of API requests, a harvester:

1. Terminates warcprox.
2. Sends a new harvest request message to the web harvester containing the extracted links.
3. Moves the WARC file created by warcprox and sends a message announcing its creation to the messaging queue. The WARC created message is a useful hook for other services. For example, SFM UI listens for WARC created messages and creates database records to keep track of the WARC files.
4. Sends a harvest response message. The harvest response message contains information on the outcome of the harvest such as its success/failure, basic statistics, and possible updates to seeds. SFM UI uses the harvest response message to update the database records that keep track of harvests.

The web harvester is implemented differently than the social media harvesters. The web harvester wraps Internet Archive's Heritrix, which takes care of retrieving the web resources and writing to WARC files [9]. When capturing a web page, Heritrix is configured to retrieve only that page and its dependencies; it does not crawl websites.

In our initial usage of SFM, we have found that harvesting the web resources for social media data takes significantly longer than harvesting the social media data itself; further, the WARC files for the related web resources require much more storage than the social media data. Neither of these is unexpected, but will need additional engineering to support scaling of harvesting-related web resources. Complementary alternatives include only harvesting the top referenced related web resources or submitting the URLs to the Internet Archive for harvesting.

7.3 Exporting and processing

Exporting and processing of social media data depend on the ability to extract the posts from the WARC files. To support this, we have developed separate WARC iteration libraries for each social media platform, but they all follow the same steps:

1. Use Internet Archive's WARC library to read the WARC file, extract WARC records, and parse the WARC record headers [73].
2. If, based on the URL contained in the WARC record header, a WARC response record contains data to be exported, load the WARC record payload into the requests library. The requests library is a common Python library for handling HTTP. In this case, the request library will handle the content encoding and chunk transfer encoding and parse the JSON record.
3. Finally, extract posts from the JSON record and return one-by-one. The structure of the JSON record will vary by social media platform, so extracting posts from the JSON record is platform-specific.

Thus, the result of iteration through an SFM WARC looks similar to the result of executing "cat" on a line-oriented JSON file.

---

[16] https://dev.twitter.com/rest/reference/get/statuses/user_timeline.

Building on the foundation of the WARC iteration libraries, SFM provides a fair bit of flexibility in exporting and processing social media datasets for users. First, a user can request an export of a collection using SFM UI. Users can select to limit the export by seed, harvest date, and/or item date. The post and a subset of its metadata can be exported to a variety of formats including Excel, CSV, and JSON; alternatively, just the post IDs can be exported to a text file[17] or in their entirety to JSON.

When a user requests an export, an export request message is sent to the message queue. The export request messages contain the information needed by an exporter to perform an export. Upon receiving the export request message, the exporter will do the following:

1. Use SFM UI's API to determine which WARC files might contain social media data to export. Right now this is based on collections and harvest dates, but there is potential for providing better limiting of WARC files for greater efficiency.
2. Iterate over the posts contained in each of the WARC files. Depending on the export request, the posts may be further filtered. Each selected post is written to an export file in the appropriate format. The result is a single file for the export.

When completed, the exporter sends an export response message. SFM UI uses the export response message to update the database records that keep track of exports and notifies the user by email.

To support other researcher usage, e.g., piping data directly into an analytics tool or performing more advanced filtering, exports can be performed from the command line. A user can invoke a WARC iteration library directly on a WARC file, getting a list of posts. To help a user determine which WARC files to export from, a command line utility (find_warcs.py) is provided that will query SFM UI's API and return a list of WARC files. In addition, to assist with exporting from the command line, a Docker container is available that is preconfigured with various SFM utilities, as well as some other useful tools (e.g., jq [74], Twarc [71], JWAT Tools [75], and warctools [76]).

7.4 Discovery

As mentioned, our philosophy is to defer to the researcher for the selection of appropriate analytic tools. However, for the purposes of monitoring and adjusting the targets of ongoing social media collections, some rudimentary analysis of the collected social media data is useful. To this end, SFM includes a Docker container with an instance of the ELK (Elasticsearch, Logstash, Kibana) stack[18] that has been customized for exploring social media data. The ELK stack is a general-purpose framework for exploring data. It provides support for loading, querying, analyzing, and visualizing. In version 1.0, SFM's ELK instance handles Twitter and Weibo data. Figure 8 is an example of a visualization.

7.5 Playback experiment

The one aspect of the technical approach that is not implemented in the current version of SFM is the playback of social media content and related web resources using wayback software. However, we have performed some experimentation to prove the feasibility of this approach.

First, we loaded WARC files with social media content and related web resources into pyWB, an instance of wayback software written by Ilya Kreymer [14]. This allowed finding calls to social media APIs by URL and date, as shown in Fig. 9.

PyWB resolves the links contained in the social media data so that they resolve to archived web resources, as shown in Fig. 10.

We then wrote a web application (using AngularJS) that would retrieve the social media data from PyWB and render it in a human-viewable form (Fig. 11).

Within the scope of the NHPRC grant, this form of Web-like access to social media data will not be further developed. However, we hope to find the opportunity to do so in the future, and we welcome ideas and potential collaborators in this area.

8 Conclusion and future work

While this reenvisioned approach to social media collecting is exploratory, thus far it has been borne out by the progress we have made in rewriting SFM. Along with further software development and experience, we hope to benefit from the scrutiny of researchers, archivists, the web archiving community, and other potential users of social media collections. With the release of version 1.0 of SFM, we have been working to provide test instances to willing representatives of these groups to spark adoption and gather feedback.

As outlined in our development roadmap,[19] there is still substantial work to be performed on SFM, as follows:

- Better operationalization, including improved monitoring and logging.

---

[17] For Twitter, this is commonly referred to as "dehydration" and is useful because it allows exchanging datasets within the constraints of Twitter's terms of service.
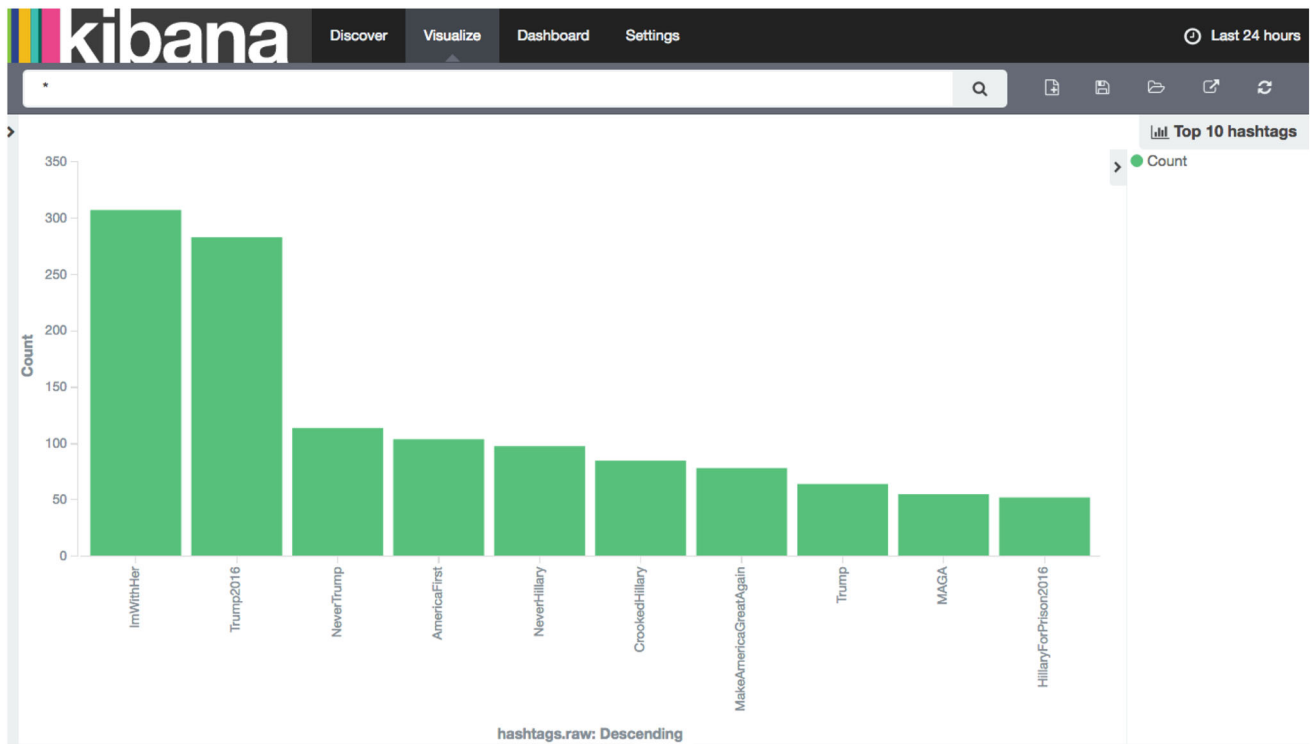
[18] https://www.elastic.co/.

[19] http://gwu-libraries.github.io/sfm-ui/about/roadmap.

Fig. 8 Visualization with ELK

Fig. 9 Calls to flickr API as returned by pyWB

```
- {
    created_at: "Sun Apr 03 00:09:52 +0000 2016",
    id: 716417396140064800,
    id_str: "716417396140064768",
    text: "Things you never think you'll need to tweet: Please don't leave your banana on a book shelf. https://t.co/V3uU0VNPI
  - entities: {
        hashtags: [ ],
        symbols: [ ],
        user_mentions: [ ],
      - urls: [
          - {
                url: "http://192.168.99.100:8082/8f8f78a451084eb1ab48d9521c03cde9/20160403221952/https://t.co/V3uU0VNPFF",
                expanded_url: "http://192.168.99.100:8082/8f8f78a451084eb1ab48d9521c03cde9/20160403221952/https://twitter.com/c
                display_url: "twitter.com/desireehalpern…",
              - indices: [
                    94,
                    117
                ]
            }
        ]
    },
    truncated: false,
  - metadata: {
        iso_language_code: "en",
        result type: "recent"
```

Fig. 10 Social media data with resolved links

Fig. 11 Rendering of a tweet

- Continued refinement of usability.
- Exposing the provenance metadata as required by researchers.

Though not yet on the development roadmap, some additional work beckons such as Web-like access to social media data as described above. In addition, there are other techniques from web archiving that might be considered, such as using CDX indexes for WARC files to speed access to social media data.

Beyond development, one area where we see great promise in aligning social media collecting with web archiving is the opportunity to jointly engage in a conversation with social media researchers and archivists about how to support quality research and robust archives. As historians and other researchers use web archives, their experiences contribute to future requirements. We hope to see this discussion lead to establishing best practices and standards for social media data collection. Further, this understanding will inform future work on some of the more speculative areas discussed above,

such as approaches to access and discovery of collected content, data visualization, and other forms of analysis of social media and web archives.

References

1. GW Libraries: gwu-libraries/social-feed-manager (2012). https://github.com/gwu-libraries/social-feed-manager. Accessed 10 Feb 2016
2. GW Libraries: Welcome to Social Feed Manager! (2015). http://social-feed-manager.readthedocs.org/en/latest/. Accessed 12 Feb 2016
3. Chudnov, D., Kerchner, D., Sharma, A., Wrubel, L.: Technical challenges in developing software to collect twitter data. Code4lib J. (2014) http://journal.code4lib.org/articles/10097. Accessed 10 Feb 2016
4. Hayes, D., Lawless, J.L.: Women on the run: gender, media, and political campaigns in a polarized Era. Cambridge University Press, Cambridge (2016). http://books.google.com/books/about/Women_on_the_Run.html?hl=&id=fXNNDAAAQBAJ. Accessed 10 Feb 2016
5. GW Libraries: gwu-libraries/sfm-ui (2015). https://github.com/gwu-libraries/sfm-ui. Accessed 10 Feb 2016
6. GW Libraries: Social Feed Manager (SFM) documentation (2015). http://sfm.readthedocs.org/en/latest/. Accessed 12 Feb 2016
7. International Internet Preservation Consortium: About IIPC (2012). http://netpreserve.org/about-us. Accessed 10 Feb 2016
8. International Internet Preservation Consortium: About archiving (2012). http://netpreserve.org/web-archiving/about-archiving. Accessed 10 Feb 2016
9. Jack, P., Levitt, N.: Heritrix (2014). https://webarchive.jira.com/wiki/display/Heritrix. Accessed 10 Feb 2016
10. Kreymer, I.: Webrecorder/webrecorder (2015). https://github.com/webrecorder/webrecorder. Accessed 11 Feb 2016
11. Internet Archive: Internetarchive/warcprox (2012). https://github.com/internetarchive/warcprox. Accessed 11 Feb 2016
12. International Internet Preservation Consortium: The WARC format (2015). http://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/. Accessed 10 Feb 2016
13. International Internet Preservation Consortium: iipc/openwayback (2013). https://github.com/iipc/openwayback. Accessed 11 Feb 2016
14. Kreymer, I.: ikreymer/pywb (2013). https://github.com/ikreymer/pywb. Accessed 11 Feb 2016
15. Thomson, S.D.: Preserving social media (2016). doi:10.7207/twr16-01. http://www.dpconline.org/component/docman/doc_download/1486-twr16-01. Accessed 10 Feb 2016
16. Bercovici, J.: Who coined "Social Media"? Web pioneers compete for credit. Forbes. (2010). http://www.forbes.com/sites/jeffbercovici/2010/12/09/who-coined-social-media-web-pioneers-compete-for-credit/. Accessed 10 Feb 2016
17. Espley, S., Carpentier, F., Pop, R., Medjkoune, L.: Collect, preserve, access: applying the governing principles of the national archives UK government web archive to social media content. Alexandria **25**, 31–50 (2014). doi:10.7227/ALX.0019. http://openurl.ingenta.com/content/xref?genre=article&issn=0955-7490&volume=25&issue=1&spage=31. Accessed 10 Feb 2016
18. Bragg, M., Eubank, K., Ricker, J.: Preserving Web 2.0. Presented at: Best practices exchange (2009) https://webarchive.jira.com/wiki/download/attachments/5734676/BPE_web2_partner+meeting.ppt?version=1&modificationDate=1257454424180. Accessed 10 Feb 2016
19. Ricker, J.: A flickr of Hope: harvesting social networking sites with archive-it. Presented at: NDIIPP partners meeting (2010). http://digitalpreservation.ncdcr.gov/asgii/presentations/ndiipp2010.pdf. Accessed 10 Feb 2016
20. Ricker, J.: Archiving social media sites in North Carolina. Presented at: Best practices exchange (2010). http://digitalpreservation.ncdcr.gov/asgii/presentations/bpe2010.pdf. Accessed 10 Feb 2016
21. Trent, R., Kenney, K.: Social Media Archiving in State Government. Presented at: Tri-State archivists meeting (2013). http://digitalpreservation.ncdcr.gov/asgii/presentations/snca_2013_socialmedia.pdf. Accessed 10 Feb 2016
22. McNealy, J.E.: The privacy implications of digital preservation: Social media archives and the social networks theory of privacy. Elon Univ. Law Rev. **3**, 133–160 (2010). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2027036. Accessed 10 Feb 2016
23. Miao, T.A.: Access denied: how social media accounts fall outside the scope of intellectual property law and into the realm of the computer fraud and abuse act. Fordham Intell. Prop. Med. Ent. LJ **23**, 1017 (2012). http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/frdipm23&section=32. Accessed 10 Feb 2016
24. Moyer, M.W.: Twitter opens its cage. Sci. Am. **310**, 16 (2014). http://www.ncbi.nlm.nih.gov/pubmed/25004563. Accessed 10 Feb 2016
25. NDSA Content Working Group: Web Archiving Survey Report. National Digital Stewardship Alliance (2012). http://www.digitalpreservation.gov/ndsa/working_groups/documents/ndsa_web_archiving_survey_report_2012.pdf. Accessed 10 Feb 2016
26. Bowers, K., Dolan-Mescal, A., Donovan, L., et al.: Occupy archives panel. Presented at: Annual Society of American Archivists Meeting (2013). http://archives2013.sched.org/event/14m52JH/session-303-occupy-archives. Accessed 10 Feb 2016
27. King, L.: Emory digital scholars archive occupy wall street Tweets. Emory Rep. (2012). http://news.emory.edu/stories/2012/09/er_occupy_wall_street_tweets_archive/campus.html. Accessed 10 Feb 2016
28. Del Signore, J.: Museums Archiving Occupy Wall Street: Historical Preservation Or "Taxpayer-Funded Hoarding"? Gothamist (2011). http://gothamist.com/2011/12/26/occupy_wall_street_the_museum_exhib.php. Accessed 10 Feb 2016
29. Chitturi, K., Yang, S.: Real-time archiving of spontaneous events (Use-Case; Hurricane Sandy) and visualizing disaster phases appearing in Tweets. Presented at: Archive-it partner meeting at Best practices exchange. (2012). https://webarchive.jira.com/wiki/download/attachments/40075274/Real-%C2%AD%E2%80%90%26me%20Archiving%20of%20Spontaneous%20Events%20%28Use-%C2%AD%E2%80%90Case%20-%20Hurricane%20Sandy%29.pdf. Accessed 10 Feb 2016
30. Gueguen, G.: Capturing the Zeitgeist. (2012). http://www.slideshare.net/guegueng/capturing-the-zeitgeist. Accessed 10 Feb 2016
31. National Archives and Record Administration: Best practices for social media capture. National Archives and Record Administration (2013). http://www.archives.gov/records-mgmt/resources/socialmediacapture.pdf. Accessed 10 Feb 2016
32. Trent, R.: Social media archive BETA is live! The G.S. 132 Files (2012). https://ncrecords.wordpress.com/2012/12/04/social-media-archive-beta-is-live/. Accessed 10 Feb 2016
33. Emory Libraries: emory-libraries/Twap (2011). https://github.com/emory-libraries/Twap. Accessed 11 Feb 2016

34. North Carolina State University Libraries: NCSU-Libraries/lentil (2013). https://github.com/NCSU-Libraries/lentil. Accessed 11 Feb 2016

35. Thomson, S.D., Kilbride, W.: Preserving social media: the problem of access. New Rev. Inf. Netw. **20**, 261–275 (2015). doi:10.1080/13614576.2015.1114842

36. Pennock, M.: Web-archiving (2013). doi:10.7207/twr13-01

37. Bailey, J., Grotke, A., Hanna, K., et al.: Web archiving in the United States: a 2013 survey. National Digital Steward-ship Alliance (2014). http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_USWebArchivingSurvey_2013.pdf. Accessed 10 Feb 2016

38. Boyd, D., Crawford, K.: Critical questions for big data. Inf. Commun. Soc. **15**, 662–679 (2012). doi:10.1080/1369118X.2012.678878

39. Bruns, A.: Faster than the speed of print: reconciling "big data" social media analysis and academic scholarship. First Monday (2013). doi:10.5210/fm.v18i10.4879. http://journals.uic.edu/ojs/index.php/fm/article/view/4879. Accessed 20 July 2016

40. Tufekci, Z.: Big questions for social media big data: representative-ness, validity and other methodological pitfalls. arXiv:1403.7400

41. Hajtnik, T., Uglešić, K., Živkovič, A.: Acquisition and preserva-tion of authentic information in a digital age. Publ. Relat. Rev. **41**, 264–271 (2015). doi:10.1016/j.pubrev.2014.12.001. http://www.sciencedirect.com/science/article/pii/S0363811114001945. Accessed 10 Feb 2016

42. Eltgrowth, D.R.: Best evidence and the Wayback machine: toward a workable authentication standard for archived Internet evidence. Fordham Law Rev. **78**, 181 (2009). http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/flr78&section=8. Accessed 10 Feb 2016

43. AIIM: AIIM TR31-2004, Legal acceptance of records produced by information technology systems (2004). http://www.aiim.org/Resources/Standards/AIIM_TR_31. Accessed 20 July 2016

44. State Archives of North Carolina: Guidelines for managing trust-worthy digital public records (2000). http://archives.ncdcr.gov/Portals/3/PDF/guidelines/guidelines_for_digital_public_records.pdf. Accessed 10 Feb 2016

45. Markham, A., Buchanan, E., Committee, A.E.W. Others: Eth-ical decision-making and Internet research: Version 2.0. Asso-ciation of Internet Researchers (2012). http://www.uwstout.edu/ethicscenter/upload/aoirethicsprintablecopy.pdf. Accessed 10 Feb 2016

46. Leetaru, K.: Are research ethics obsolete in the Era of big data? Forbes (2016). http://www.forbes.com/sites/kalevleetaru/2016/06/17/are-research-ethics-obsolete-in-the-era-of-big-data/. Accessed 20 July 2016

47. Council for Big Data, Ethics, and Society (2016). http://bdes.datasociety.net/. Accessed 20 July 2016

48. Summers, E.: Introducing documenting the now—documenting DocNow. Medium (2016). https://news.docnow.io/introducing-documenting-the-now-416874c07e0. Accessed 25 July 2016

49. Townsend, L., Wallace, C.: Social media research: a guide to ethics. The University of Aberdeen. http://www.dotrural.ac.uk/socialmediaresearchethics.pdf. Accessed 10 Feb 2016

50. Milligan, I., Webster, P.: The Web archive bibliography. Web archives for historians (2014). https://webarchivehistorians.org/the-web-archive-bibliography/. Accessed 22 July 2016

51. Milligan, I.: Finding community in the Ruins of GeoCities: dis-tantly reading a web archive. Bull. IEEE Tech. Commit. Dig. Lib. (2015). http://www.ieee-tcdl.org/Bulletin/v11n2/papers/milligan.pdf. Accessed 10 Feb 2016

52. Milligan, I.: Lost in the infinite archive: the promise and pitfalls of web archives. Int. J. Hum. Arts Comput. **10**, 78–94 (2016). doi:10.3366/ijhac.2016.0161

53. Webster, P.: Why historians should care about web archiving. Webstory: Peter Webster's blog. (2012). https://peterwebster.me/2012/10/08/why-historians-should-care-about-web-archiving/. Accessed 14 July 2016

54. Statista (2016) Twitter: number of monthly active users 2015. Statista. http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/. Accessed 11 Feb 2016

55. Summers, E.: URLs in Tweets Mentioning Ferguson: August 10–27, 2014 (2014). https://edsu.github.io/ferguson-urls/index.html. Accessed 10 Feb 2016

56. Baumann, R.: Archiving Video from #Ferguson: on Archivy. Medium. (2015) https://medium.com/on-archivy/archiving-video-from-ferguson-504e95859756. Accessed 10 Feb 2016

57. Milligan, I., Ruest, N., Lin, J.: The Gatekeepers vs. the Masses. Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. (2016). doi:10.1145/2910896.2910913

58. Consultative Committee for Space Data Systems: Reference model for an Open Archival Information System (OAIS). CCSDS Secretariat, Washington, DC (2012). http://public.ccsds.org/publications/archive/650x0m2.pdf. Accessed 10 Feb 2016

59. Commission on Preservation and Access, Research Libraries Group, Task Force on Digital Archiving: Preserving digi-tal information: report of the task force on archiving of digital information (1996). https://books.google.com/books?id=T9YmrgEACAAJ. Accessed 10 Feb 2016

60. Provenance Working Group: PROV-overview (2013). https://www.w3.org/TR/prov-overview/. Accessed 6 June 2016

61. Kerchner, D., Littman, J., Peterson, C. et al.: The Provenance of a Tweet (2016). https://scholarspace.library.gwu.edu/downloads/h128nd689. Accessed 20 July 2016

62. Internet Archive: internetarchive/brozzler. GitHub. https://github.com/internetarchive/brozzler. Accessed 14 July 2016

63. Littman, J.: Social media harvesting techniques. GW Libraries (2015). https://library.gwu.edu/scholarly-technology-group/posts/social-media-harvesting-techniques. Accessed 10 Feb 2016

64. Foo, C.: chfoo/wpull (2013). https://github.com/chfoo/wpull. Accessed 10 Feb 2016

65. Van de Sompel, H., Nelson, M., Sanderson, R.: HTTP framework for time-based access to resource states–Memento (2013). https://tools.ietf.org/rfc/rfc7089.txt. Accessed 10 Feb 2016

66. Wrubel, L.: Announcing SFM Version 1.0. Social Feed Man-ager (2016). http://gwu-libraries.github.io/sfm-ui/posts/2016-06-20-releasing-1-0. Accessed 15 July 2016

67. Twitter, Inc. The Streaming APIs. https://dev.twitter.com/streaming/overview. Accessed 20 July 2016

68. REST APIs. Twitter Developers. https://dev.twitter.com/rest/public. Accessed 20 July 2016

69. Flickr.: Flickr services (2005). https://www.flickr.com/services/api/. Accessed 20 July 2016

70. Weibo Corporation: Weibo API (2012). http://open.weibo.com/wiki/API%E6%96%87%E6%A1%A3/en. Accessed 20 July 2016

71. Summers, E.: edsu/twarc (2013). doi:10.5281/zenodo.17385. https://github.com/edsu/twarc. Accessed 10 Feb 2016

72. Stüvel, S.A.: sybrenstuvel/flickrapi (2013). https://github.com/sybrenstuvel/flickrapi. Accessed 18 Oct 2016

73. Internet Archive: internetarchive/warc. GitHub. https://github.com/internetarchive/warc. Accessed 15 July 2016

74. Dolan, S.: stedolan/jq (2012). https://github.com/stedolan/jq. Accessed 18 Oct 2016

75. Clarke, N.: JWAT-tools (2012). https://sbforge.org/display/JWAT/JWAT-Tools. Accessed 18 Oct 2016

76. Internet Archive: internetarchive/warctools (2010). https://github.com/internetarchive/warctools. Accessed 18 Oct 2016