

# Social Feed Manager

## Guide for Building Social Media Archives

Christopher J. Prom  
University of Illinois at Urbana-Champaign  
June 7, 2017

## Table of Contents

<b>Introduction and Project Background</b>	<b>4</b>
Need for Social Media Capture and Role Played by SFM	5
About this Report	6
SFM's Position vis-à-vis Other Tools	8
SFM's Advantages	9
SFM's Limitations	11
<b>Getting Started with Social Feed Manager</b>	<b>12</b>
Learning About the Project	12
Implementation Options	13
Running Locally	13
Running on a Server	14
Understanding SFM as a Technology	15
Basic Technical Structure	15
Data Model Overview and Implications for Preservation	17
Seeds and Collection Types	18
Practical Implications	19
Developing Collection Sets, Collections, Seeds, and Harvests	23
Twitter User Timeline	24
Twitter Search	26
Twitter Sample	27
Twitter Filter	27
Flickr User	29
Weibo Timeline and Weibo Search	30
Tumblr Blog Posts	31
Web Resources	32
Shaping Local Services	33
<b>Use cases</b>	<b>36</b>

GW Libraries User Stories	36
Institution-Led Collecting	37
Facilitating Records and Information Compliance	37
Capturing Institution-Related Materials	38
Documenting Events	40
Collecting Topical or Subject-Based Records	40
Research Consultation Service	41
Researcher Collects, Uses, and Discards	42
Researcher Collects and Library Preserves	42
Library/Archives Creates and Preserves	42
<b>Preservation of and Access to Social Media Records</b>	<b>44</b>
Establishing a Policy Basis for Access	44
Export Options and Scenarios	45
Exporting Data through the User Interface	46
Copying Datasets from the Server	48
Packaging Recommendations and Descriptive Metadata	48
Export-Based Packaging	49
Collection Set or Collection-Based Packaging	50
Combination Packaging	52
Recording Descriptive Metadata	53
Accruals and Pruning	54
<b>Community Recommendations</b>	<b>55</b>
For Developers	55
Enhance Capture Capabilities	55
Refine Export and Packaging Options	56
For Libraries and Archives	57
Establish Social Media Collecting as Core Archival Activity	57
Advocate for Enhanced Preservation and Access Rights	57
For the Community and Funding Partners	58
<b>Conclusion</b>	<b>60</b>
<b>Appendices</b>	<b>61</b>
Appendix 1. Reference List/Further Readings	61
Appendix 2. Tools and Services for Archiving Social Media	64
ArchiveIT	64
ArchiveSocial	64
Hydrator	64
Lentil	65

Smarsh Archiving Platform	66
Twarc	66
Twitter Archiving Google Sheet (TAGS)	66
Twitter Archive	67
Facebook Account Download	67
WebRecorder	68

## Introduction and Project Background

In their daily lives, many Americans access or post information using social media services such as Twitter, Instagram, Tumblr, and of course Facebook.<sup>1</sup> In addition, many businesses, politicians, celebrities, government agencies, media companies, and nonprofit organizations use social media to publicize their activities, to share information with members of the public, or to influence public opinion.

This activity is part of our cultural landscape, but it takes on new meaning when viewed from an archivist's perspective. Whenever a person or organization posts something using a social media service, they create a record, a record that documents a personal or corporate activity, event, communication, thought, or opinion. As such, social media posts leave real-time traces in the complex Internet ecosystem. A single post may reference a blog article, newspaper story, press release, facebook posting, or video; posts are reposted, remixed, and reused; and people send posts to each other using different platforms or messaging applications. And not only do social media records (both individual posts and aggregations of post) reflect what people think, do, or feel, but they influence public opinions and perceptions of events. They shape history.

To see the truth of this, one need look no further than accounts of prominent individuals, including our former and especially our current President. Even the accounts of everyday citizens offer grist for the mills of those seeking to document or understand life in the 21st century; but days, weeks, months or years after a post was shared, it can be difficult for people to understand the original context and significance of social media postings, or even to find posts that relate to a topic of interest.<sup>2</sup>

Social Feed Manager (SFM) was developed as one method to address this problem. It makes it possible for scholars, citizens, and archivists to capture, preserve, and study the constantly evolving records that are generated when people and corporate bodies use social media services.

This report provides basic advice regarding SFM and outlines a few use cases, so that people working in a variety of curatorial settings can implement and use SFM to collect social media records and then to make those records accessible to future students, scholars, and members of the public. In addition, it recommends specific strategies that archivists can use to export and package social media data, taking account of strategies that collecting organizations typically employ when accessioning, describing, preserving, and storing archival materials. Finally, the report speculates a bit as to how SFM might be enhanced to make future versions of the tool even more useful, given work done on the project to date.

---

<sup>1</sup> Greenwood, Perrin, and Duggan, "Social Media Update 2016."

<sup>2</sup> Thomson, *Preserving Social Media* summarizes the importance of social media as a documentary resource and reviews the case for its inclusion in libraries and archives.

## Need for Social Media Capture and Role Played by SFM

In 1974, Theodor Nelson wrote "Everything is deeply intertwined. [...] There are no "subjects" at all; there is only all knowledge, since the cross-connections among the myriad topics of this world simply cannot be divided up neatly."<sup>3</sup> Put more simply, our world is dominated by relationships between objects, ideas, people, places, and activities. Social media records reflect this chaos, mirroring reality and pushing archivists to embrace new practices.

Let us imagine that thirty years from now we want to embody the tweets shown in figure one. We want to give people ability to understand them as points in the multi-dimensional and time-bounded space. To fully do this, we would need to capture their relationships to each other, to other objects that they reference, and maybe to other objects that reference them. And this leaves aside other implicit knowledge that might be valuable to understanding the tweets, such as the meaning of the hashtags #moab and #maga.



**Figure One: Sample Tweet from Donald Trump Jr. and reply by James Fallows**

Capturing everything that is potentially related to these tweets is clearly an impossible task, not to mention an undesirable and unwise one. SFM allows us to capture a particular representation of a social media record and the ecosystem it inhabited, a providing a point-in-time snapshot. SFM also captures metadata that is not readily apparent but which may become essential in the future use or interpretation of tweets like these. For example, archivists might use SFM to capture the following information about the Trump and Fallows tweets:

- The JSON version of the tweets, which includes associated metadata like screen names. It can capture the user profiles in place when the tweet was posted as well as provenance metadata about the tweet itself (such as the sending device, date and time of posting, date and time of captured)

<sup>3</sup> Nelson, *Computer Lib; Dream Machines*, page DM45.

- Copies of specific resources referenced by that tweet, such as a profile images, embedded videos, and linked webpages.
- Retweets, replies, and other mentions of the tweets.

How does SFM work? Specifically, it sends requests to Twitter’s Application Programming Interface (API), then stores the responses: a machine readable version of the tweet in JSON format. If the user instructs it to, SFM then uses the web harvester Heritrix to request related resources such as embedded images, profile images, and referenced webpages.

### **What are JSON and APIs?**

JSON stands for JavaScript Object Notation. It is a text-based way to store and structure information. Since JSON objects can be easily parsed and displayed using Javascript, modern web servers and clients use it as a common exchange format. JSON files can be ‘flat’ (that is represent a simple object), or they can contain nested data in a simple hierarchy, like an XML document, but simpler.

JSON provides programmers several benefits, including the ability to dynamically load content as you scroll through a page, to re-populate a form without reloading the entire page, or to otherwise change the display based on some user action.

Servers typically send JSON to a client computer after the client sends a request to an Application Programming Interface (API), that is, a set of tools and commands that a developer can use to request information. The Social Feed Manager uses the APIs from Twitter, Tumblr, Flickr and Sina Weibo to request data from those services. It then copies them to SFM’s data store.

When such information is captured for dozens, hundreds, thousands, or millions of social media records, they generate a store of related data, one that can be queried in ways limited only by capabilities of software and the imaginations of present and future researchers. In other words, SFM allows archivists and researchers the opportunity to develop and capture data collections—whether those are small and closely refined collections gathered to meet an immediate research need or massive, widely ranging dataset amassed for longer-term preservation.

### **About this Report**

The data collection aspects of SFM are reflected in its history: SFM originated in faculty-led collecting efforts at George Washington University Libraries (GW Libraries). GW Libraries developed it to help scholars collect, study, and report upon specific datasets, such as a collection of tweets or tumblr posts related to a particular topic or hashtag.<sup>4</sup> The tool should be of great interest to archivists working in a range of a settings, including academic and

---

<sup>4</sup> Littman, Justin, et al, “API-Based Social Media Collecting.”

government archives, businesses, and nonprofit organizations, as well as to the end users of archives and libraries.

Among SFM's potential users, archivists and other information professionals have a particular charge: to preserve information for future use and access, not just for its current utility. And even those who use tools like Social Feed Manager to collect records for a current project (such as a linguistics student analyzing the semiotics of hashtags) may wish to partner with archivists and librarians, who can breathe future life and value into the collections they assembled for their own needs. For these reasons, archivists and librarians compose the primary audience for this report.

Of course, SFM will interest many people whom libraries or archives serve: scholars, students, and members of the public. For example, these individuals may wish to seed a particular data-capture project into the tool, then export the records and analyze them using topic modeling, data mining, visualization software, or another method.<sup>5</sup> Or they may wish to analyze a previously captured collection. Such people are a secondary audience for this report, if only since they will most likely use the tool in consultation with an archivist or librarian, whose institution maintains an instance of the tool. The focus of this report is on helping organizations implement SFM, so they can use it themselves to begin building social media archives, while also helping others to use the tool for their research.

The report is one output from a project to assess and provide feedback regarding Social Feed Manager. Under the terms of a grant from the National Historical Publications and Records Commission, George Washington University conducted a three-year project to explore innovative methods to facilitate the preservation and use of social media collections by developing a viable tool to capture social media records and metadata. GW Libraries also developed use cases and policy studies concerning the long-term preservation and access that might be afforded to those captured records.

As consultant, I assessed the tool in light of these project goals, offered some recommended enhancements, and developed this report. The work herein is based on extensive testing and use of the tool both on a server provided by project staff, as well as on local instances running on desktop computers and a server at the University of Illinois, where SFM is currently being configured and assessed for production implementation. We plan to replace prior social media capture projects, which used the Twitter Archiving Google Sheet, with projects centered instead on SFM.<sup>6</sup>

---

<sup>5</sup> SFM includes limited internal data visualization capacities. The current work focus was to improve and refine its data collection and packaging capacities, recognising that different people will analyze records in different ways, using different tools.

<sup>6</sup> Hawksey, "TAGS Website."

## SFM's Position vis-à-vis Other Tools

Social media records are but one of many types of records that an archives or library may wish to capture, but they are one that requires specialized tools. Sarah Day Thomson describes the following basic approaches, each of which can be used to capture or 'archive' social media data: (1) querying a platform's Application Programming Interfaces (APIs); (2) purchasing data from a reseller; (3) contracting with a third-party service; and (4) downloading data via a platform-provided self-archiving tool.<sup>7</sup> She discusses the pros and cons of API collecting and cites a few very early projects, in which European archives or libraries developed local tools to collect social media records.<sup>8</sup> While Thomson listed a few open source tools in an appendix, they were not highly enough developed to warrant further discussion at that time. In addition to the strategies that she mentions, I would add a fifth category: 'traditional' or 'enhanced' web archiving, using tools like wget, the Internet Archive (or the paid service Archive-It) or Webrecorder.<sup>9</sup>

At the time of writing (May 2017), social media archiving is still at an earlier phase in its development than comparable areas like web archiving or email archiving, and the challenges and opportunities that Thomson lays out have not fundamentally changed. For this reason, any archives considering whether to capture social media should start by reading her report, while recognizing that the tools have evolved somewhat since 2016.

Figure two compares the five approaches and describes some current tools. This figure and the tool descriptions in Appendix B can be used as a starting point, to help your archives chose the tools that best suit your goals

**Figure Two: Social Media Archives Tools and Services**

Category	Example Tools or Services	Advantages	Disadvantages
Open-source API-based collectors	Twarc; hydrator; Twitter Archiving Google Sheet;	Flexibility and extensibility; Data under local control	Requires local implementation support Data capture and use

<sup>7</sup> Thomson, "Preserving Social Media," 9-14.

<sup>8</sup> Ibid., 28-32.

<sup>9</sup> What I am calling traditional tools include wget (<https://www.gnu.org/software/wget/>) and Herertrix (<https://webarchive.jira.com/wiki/display/Heritrix>), which is used by the Internet Archive, Archive-It, and other web archiving services. Such tools issue a standard http request then preserve whatever is returned to them, typically (but not exclusively) the components of a webpage that are served back through a http response (such as html files, stylesheets, images, javascript). What I am calling 'enhanced' tools include the webrecorder ( <https://webrecorder.io/>; <https://github.com/webrecorder/webrecorder>). The webrecorder captures http responses as a user browses the web, interacts with forms and loads dynamic content. After the recording session has been finalized, the user can bookmark particular pages and describe the archives using markdown.



	Social Feed Manager; Lentil; DocNow <sup>10</sup>	and management.	bound by API terms of service.
Data resellers	Gnip; DiscoverText	Access to historical data and additional metadata	High cost; requires additional development for capture, preservation, and access
Third-party service providers	Archive Social; Smarsh	Ease of implementation	Cost unpredictable; Same API limits targeted at records management and legal compliance; Potential data lock in; Requires preservation planning;
Platform self-archiving	Twitter Archive; Facebook	Ease of implementation Access to restricted spaces (i.e. Facebook)	Requires access to user credentials; data limitations; does not include embedded or linked resources
Web Archiving	Wget; Heritrix; Archive-It; Webrecorder	Allows fine control; webrecorder shows interactive nature of dynamic content)	Not tailored to social media; Archive-iT and Webrecorder available as hosted services, but neither is specifically tailored to social media.

## SFM's Advantages

Institutions can select from many options but should consider project objectives, preservation needs, and resource requirements before simply picking a tool. That said, SFM occupies a particular and important place in the suite of potential tools that archivists and others can use to capture and use social media records, and it bears a particularly close look. There are four reasons for this.

First, SFM is an open source tool that should be accessible to most any archivist. Unlike other open source tools, it does not require extensive technical knowledge to implement and manage, either in a desktop or server scenario. As described in more detail below, it can be installed and run on a desktop computer for testing (although best implemented on a dedicated server).

<sup>10</sup> DocNow maintains the hydrator and twarc software, as well as providing a distribution method for tweet ID datasets. As the project evolves may release some additional tools.

Unlike other tools, it can be provided as a service to end users, who can manage the collection of social media datasets with a simple and intuitive web interface. While SFM uses many software libraries and services to perform its magic, its complexities lie safely hidden from the end user. And unlike competing options, it does not require purchase or an ongoing license agreement.

Second, SFM is a very well designed, structured, and even elegant application, both in its user interface and more importantly in its backend application design and object model. While this point is discussed in more detail below, it is critical to note that SFM has been designed so that it can be expanded to harvest data from additional services, to incorporate new features, and even (in theory) to scaled across multiple servers.

Third, SFM allows archivists to directly manage that data as they see fit in their own digital repository, not as part of the service provider's 'archive,' to be exported in an incomplete way and/or only on demand. Furthermore, the tool's focus is not primarily on capturing social media data from a single stream for legal compliance reasons (as is the case with industry and government-focused tools such as Archive Social), but in capturing data from multiple streams for many potential uses.

Finally, and perhaps most significantly, SFM builds on a tried and true foundation (the web archiving approach) but adds a new wrinkle. Yes, SFM captures social media information as web archives, storing data using the WebARCchive (WARC) file format, the standard for preserving web content. However, it does not harvest the mere HTML representation of the tweet or social media post (like wget, ArchiveIT, or the webrecorder might do. Instead, it captures the structured data that is used to populate different representations of a social media post. This is a particularly important and significant advance. Why? Because archivists, curators, and librarians are charged with helping users avoid the snares that might entangle them in an infinite and intricate web of knowledge. We have the duty to provide manageable representations of archival materials and to identify the appropriate areas of knowledge and relationships that surround an archival object, so that users can come to an understanding of what the record means. That is what archivists do—they facilitate the ability of other people to make knowledge claims using archival materials. This professional function plays a fundamental role in keeping memory alive, and ideally archivists pursue it without unduly shaping that memory and while making our actions and decisions known.<sup>11</sup>

SFM excels in documenting the process by which a user and the tool collect social media records. Since it is an example of the first kind of tool described in figure two (an API-based collector), it connects to the publicly-available twitter, tumblr, flicker and Sina Weibo API's, then captures data under terms defined by those services. It harvests a much more complete range of data than other tools that use API-based capture, such as the TAGS application. Unlike other

---

<sup>11</sup> See MacNeil, "Picking Our Text;" Douglas, "Toward More Honest Description;" and Meehan, "Making the Leap from Parts to Whole."

applications, it saves the complete JSON response in its native format, not a partial or modified version of the output in some other format, such as a flat spreadsheet file. In addition, SFM incorporates Heritrix, a tool for capturing web data. It uses that tool to capture not only the data provided via the API's but (optionally) copies of resources that are embedded in or linked to a social media post.

The bottom line is simple: SFM preserves a particularly complete and well-ordered set of provenance and other metadata, which makes the records very useful for data mining, visualization, and other transformative uses. To state it directly, SFM captures more data, and of a higher quality, than other social media capture tool available to archivists.

### **SFM's Limitations**

API-based collecting such as that employed by SFM holds several advantages, but it is important to note that the access afforded by the API is limited and controlled. For example, Twitter's timeline API limits the number of tweets that can be captured to the 3,200 most recent, so archivists cannot use it to capture every tweet by a person or organization that has posted 25,000 times. If the primary project goal is to fully capture tweets from a single account or set of accounts, an archives may wish to pursue complementary strategies. It could purchase data or collaborate with account owners to download a copy of the archive directly from the platform, using an option such as twitter's 'download your archive' feature.<sup>12</sup>

Another significant limit is that SFM does not currently support the capture of data from Facebook. While it is the most popular social media platform, Facebook's API limitations make implementation of a general purpose harvester challenging at the present time. That said, the Social Feed Manager is a very powerful tool, and the way in which the tool is designed makes it extensible and sets the basis for additional development work (as noted in the community recommendations section of this report). As currently released, the Social Feed Manager supports the capture of social media records from a limited number of platforms: Twitter, Tumblr, Flickr and Sina Weibo (a Chinese social media service). SFM's architecture allows developers to add additional harvesters, so that records can be captured from other services.

But again there is an important caveat: the harvesters can only harvest what the API provides, that is to say, what the service provider shares under their terms of service. And, as noted elsewhere, social media terms of service are particularly prone to restriction and change, particularly for companies (like Facebook) that seek to keep users on their own platforms or apps developed with their developer tools, where they can monetize pageviews into revenue.<sup>13</sup> For this reason, tools such as the webrecorder, wget or other tools have a complementary role to play, as part of an integrated harvesting program.

---

<sup>12</sup> Twitter Help Center, "Downloading your Archive," <https://support.twitter.com/articles/20170160>.

<sup>13</sup> McFarlane, "How Facebook, Twitter, Social Media Make Money From You," Miglani, "How Facebook Makes Money?"

## Getting Started with Social Feed Manager

Social Feed Manager is very easy both to install and to use. By following the instructions provided in the project documentation, individuals can download, install, and test the application with about half hour's worth of work, provided that they have administrative access to a desktop computer and have basic familiarity with the commandline and systems administration. In addition, SFM can be installed on a dedicated server with a few hours of work, although this will typically require some assistance from technical staff. In my testing, I was able install the application on my laptop in less than an hour. Working with a member of our IT Staff, Jon Gorman, we had the application running on a CentOS server in less than two hours, from the time the new virtual server was launched until SFM was running and in use. In addition, my testing of the user interface indicated that relatively little training is needed to begin using the application to harvest social media records.

### Learning About the Project

George Washington University uses four main mechanisms to provide information and documentation about the Social Feed Manager:

- A project overview site: <http://gwu-libraries.github.io/sfm-ui/>, cross referenced to <http://go.gwu.edu/sfm>.<sup>14</sup>
- A documentation site: <https://sfm.readthedocs.io/en/latest/>.<sup>15</sup>
- A code repository: <https://github.com/gwu-libraries/sfm-ui>.<sup>16</sup>
- A twitter account: <https://twitter.com/socialfeedmgr>.

Before working with the Social Feed Manager, archivists or other individuals should review the project documentation and in particular the end user instructions.<sup>17</sup> In addition, the SFM project team has published many useful resources on their project site, including Collection Development Guidelines, a published article from the *International Journal on Digital Libraries*, and a working paper about provenance and metadata.<sup>18</sup>

---

<sup>14</sup> George Washington University Libraries, *Social Feed Manager Project Site*, accessed June 4, 2017, <https://gwu-libraries.github.io/sfm-ui/>.

<sup>15</sup> George Washington University Libraries, *Social Feed Manager (SFM) Documentation*, accessed June 4, 2017, <https://sfm.readthedocs.io/en/latest/>.

<sup>16</sup> George Washington University Libraries, *Sfm-UI: The [New] Social Feed Manager User Interface Application Code Repository*. Python. GWU Libraries: 2017. <https://github.com/gwu-libraries/sfm-ui>.

<sup>17</sup> See <https://sfm.readthedocs.io/en/latest/quickstart.html>

<sup>18</sup> George Washington University Libraries, "Collection Development Guidelines," <https://gwu-libraries.github.io/sfm-ui/resources/guidelines>; Littman, et al., "API-Based Social Media Collecting;" Kerchner, et al., "The Provenance of a Tweet."

Each of these is very useful in its own way and should be reviewed before working with the tool. They can help archivists make implementation decisions and develop policies that are appropriate to their particular institutional settings.

The Collection Development Guidelines are particularly helpful if an archives or special collections library wishes to use the tool to collect and preserve social media records for long term historical value, not simply to support individual's research projects. Acknowledging that the collection and use of social media records raises several potential or real ethical legal issues, the guide delineates a series of questions that can guide discussions and implementation decisions or that can inform potential discussions with administrators and counsel.

## Implementation Options

Depending on the projected uses of the tool, an archivist, librarian, or end user can use the tool in a local desktop/laptop environment or install it on a dedicated server, which is the primary intended application. When installed on a server, an organization such as an archives or library can support multiple users and uses of the tool, which would be impossible when running locally.

### *Running Locally*

The easiest way to get started with SFM is to install Docker on a local machine, such as a laptop or (preferably) desktop computer with a persistent internet connection.

Running SFM locally is **not** recommended for anything other than testing purposes. It will not be sufficient for a long term documentation project and would typically allow only one user to access the application's web interface under localhost. But it could be useful for experimentation, when access to a dedicated web server is not possible or desirable.

### What is Docker?

Docker is widely deployed and industry supported virtualization software. It allows application developers to tie many pieces of software to a single distribution packet, and it allows you to run that software on your computer, regardless of its underlying operating system.

Like many modern web applications, SFM is complex and made up of many parts. Each of these parts (for example the SFM's web harvester, its user interface, or its database) in turn relies on some underlying software of its own. Installing each of these on your computer's base operating system would be time consuming and prone to errors.

To overcome this problem, the SFM team used the docker software to segregate each of these software packages, yet put them in an overarching framework. The free Docker software you install on your local computer simply allows the various libraries to communicate with each other, regardless of the underlying operating system you are using, while also making it easier for technical staff to install and manage the application.

After testing the software on several Apple computers, I developed some instructions for installing locally.<sup>19</sup> Those who wish to install the application on locally should note that the virtualization software Docker must be installed and running on the host operating system and that no support for installing locally is provided, either by GW Libraries or by me.

Once the application is running, it will be available at <http://localhost/> under the port specified in the configuration (`.env`) file. So long as docker and the application remain running and an internet connection is live, the application will run harvests as scheduled in the user interface and will be available at localhost. The application can be stopped gracefully from the terminal at any time, with the command `docker-compose stop`.

### *Running on a Server*

If you plan to use SFM as a production application, you will want to install the application on a web server. Doing so confers several advantages including but not limited to the following factors: 1) More than one user can add and manage collection sets, collections, seeds, harvests, and exports; and users can be added to groups so they can manage shared resources; 2) Server specifications can be tuned to maximize system performance; 3) backups, exports, and server maintenance can be more easily automated; and 4) the application is structured to operate in this fashion, as an 'always on' web service.

Installing the application on a server is a very straightforward task and is a simple matter of following the instructions in the project documentation. While the task could easily be handled by a digital or systems archivist who has access to a web server, it will be best to consult with IT staff who have expertise in configuring web applications, since this will ensure that the application functions in accord with local policies, procedures, and security requirements.

Before installing SFM in a server environment, you will want to consider the follow factors, which will affect specific configuration settings and installation parameters:

1. *Base operating system*: While SFM was developed and will most often be run in a Linux environment, it will work under any operating system, assuming Docker can be installed and maintained.
2. *Docker experience*: Have IT staff previously worked with Docker applications? If not, do they require additional information to install and manage the software?
3. *Data Volume Strategy and Backups*. As noted above, SFM can store data in the docker containers or on the local file system. Regardless of which strategy is shown, the system should be placed on a system of incremental or other backups. While this will not serve as a long-term preservation strategy, it will ensure minimal security in case of a system failure.

---

<sup>19</sup> Prom, "Installing Social Feed Manager Locally."  
<http://www.archivalconnections.org/installing-social-feed-manager-locally/>

4. *System Administration*: Who will serve as the overall administrator for SFM instance, managing users, monitoring data usage, exports, and the like? While many of the administration tasks can be completed through the web UI, certain functions can be done only through the terminal. For this reason, the service manager who has overall responsibility for the content being collected will likely require either direct access to the server, or will need to work with IT staff to perform some administrative functions.
5. *Security*:
  - a. As of version 1.7, users can self register and immediately begin adding credentials, setting up harvests, etc. For these reasons, SFM must run behind a firewall, restricting access to a particular domain, IP range or other limit. The decision as to how SFM should be restricted will best be made in consultation with local IT staff, but SFM is not intended to provide a public access interface.
  - b. Passwords: As noted above, the sfmadmin account must be given a strong password. In addition, the system administrator will want to ensure that users who self register are using passwords of a sufficient complexity to avoid easy detection or brute force attacks, even if the application is running in a restricted access area.
6. *Updates*: The software provides a very clear and easy-to-follow update pathway, which is clearly described in the documentation.<sup>20</sup> Essentially this involves 1) shutting the application down, 2) backing up configuration settings, 3) downloading the current code, 4) updating configuration settings, and 5) restarting the application. In order to minimize effects on harvesting operations that may be in processor scheduled, updates should be done in close consultation with SFM's service manager.

The documentation includes some basic troubleshooting ideas, which prove helpful in case the installation goes awry or the application malfunctions. The University of Illinois initially encountered some challenges in getting the application to run smoothly on our server, issues we were eventually able resolve. If more institutions deploy SFM, they can help identify and remediate potential installation problems, and GW Libraries provides good pathways for getting help.

## **Understanding SFM as a Technology**

In order to take best advantage of SFM and in particular to preserve and make use of SFM data outside of the system, it is helpful to know how a bit about how the application works and how it stores data.

### *Basic Technical Structure*

As previously noted, SFM uses Docker to tie together the many technologies that are necessary to provide an integrated harvesting mechanism, database, user interface, and storage service.

---

<sup>20</sup> <https://sfm.readthedocs.io/en/latest/install.html#upgrading>.

Specifically, SFM depends on and uses the following technologies which are embedded in the various docker containers:

- *Apache*: The underlying web server.
- *Python and Django*: The programming languages and frameworks used by SFM development team to build the user interface, harvesters, and other elements specific to SFM and tying other elements of the system to each other.
- *Postgres*: Metadata about collection sets, collections, seeds and harvests—both that entered by human users of the system, and that generated by the harvesters—is written to a Postgres database. The database can be directly viewed and managed with the Django administration interface. This features is available only to the sfmadmin account—as well as those authorized by the sfmadmin. For authorized accounts, it can be accessed from a link under the user’s account.
- *External applications*. SFM uses some external applications or libraries in order to accomplish various tasks:
  - *Rabbit MQ*: This is a standard and widely-deployed messaging broker. It works as as a ‘middleman,’ moving various pieces of data around the system, queueing and managing harvesting and export tasks, and executing other low-level data transmission functions. While these are largely invisible to end users, RabbitMQ keeps data flowing smoothly and contributes to the overall efficiency of SFM application; it is particularly important part of the harvesters.
  - *Twarc*: twarc is a twitter harvester written and maintained by Ed Summers from the University of Maryland
  - *Heritrix*: A web harvester that is developed, used and maintained by the Internet Archive
  - *Warcprox*: also maintained by the Internet Archive, this facilitates the capture of webdata from various sources, then writes that data to the filesystem in warc format.
  - *ElasticStack*: indexes the metadata and data captured by SFM, and makes it searchable and available for some basic visualizations. This is not enabled by default, but the documentation provides clear instructions for doing so.<sup>21</sup> Once the features have been turned on, they will be available at the ports you specify in the configuration file.

Casual users and administrators need need not concern themselves with these libraries.<sup>22</sup> The fact that they were integrated into the application reflects the tool’s solid design and object structure. Each is quite well supported outside of SFM’s project team, and the code written by GW Library’s development team relies heavily upon them. Since SFM’s dependent libraries are embedded in a series of dockerfiles, the application can be deployed relatively easily. There is no need for system administrators to undertake extensive server configuration, which would be

---

<sup>21</sup> See <http://sfm.readthedocs.io/en/latest/exploring.html#exploring-social-media-data-with-elk>.

<sup>22</sup> That said, the documentation provides a very useful description and tips for using the messaging engine and writing additional harvesters, should an archives wish to collect social media from a platform other than Twitter, Tumblr, Flickr, or Weibo.

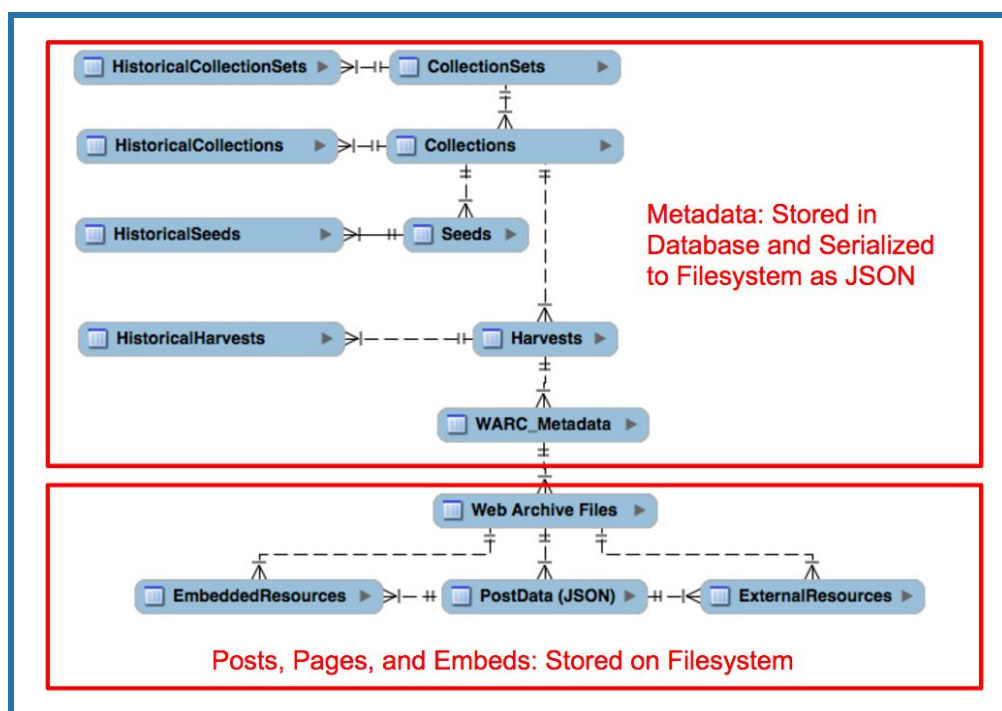


very prone to error. If staff needed to download and configure each application separately, tailoring the application to particular operating systems and versions, it would be very difficult to install, manage, and keep SFM in an operating state.

That said, the fact that SFM relies on so many external applications raises a potential maintenance issue—albeit one that can be mitigated. Whenever the code for a subsidiary library has been updated, it should also be replaced in SFM. This issue arises in most modern development projects, which tend to rely heavily on other open source libraries, and GW Libraries developed a clear means to keep the application from accruing technical debt. Overall there are fewer security and other risks to the project by using well-supported external libraries and dockerfiles, than if SFM relied on other available technologies. But the long term maintenance of SFM will need to be considered as part of a sustainability assessment, ideally one involving additional partners and institutions.

### *Data Model Overview and Implications for Preservation*

When archivists and repository staff members consider options for the long-term preservation of the data collected by SFM, they should aim to preserve both social media data and metadata about the social media records. The diagram in figure three provides a simplified overview of preservation-relevant data objects managed by the system including both objects managed in the database and in WebArchive (WARC) files:



**Figure Three: Simplified Overview of SFM's Data/Object Model<sup>23</sup>**

<sup>23</sup> Some elements of SFM's data model (such as user accounts, groups, credentials, and exports) were omitted from the diagram, since they are not immediately relevant to the following discussion.

Each of the metadata entities shown in the top boxed portion of the the entity relationship diagram corresponds to a particular data table in SFM's database.

The top-level objects managed by SFM are known as collection sets. As the notation illustrates, each collection set may contain one or more collections, and each collection can belong to only one collection set. The nature of each collection (i.e. what has been harvested or will be harvested in the future) is determined by it harvest options and by the list of seeds that have been input for each collection. Each collection can have multiple seeds, but each seed record belongs to only one collection.<sup>24</sup>

### *Seeds and Collection Types*

In SFM, a 'seed' is a harvest pathway: a defined method by which SFM queries an API and by which social media is captured and stored to SFM's data store. It is a deceptively simple concept.

SFM seeds are both similar to and different from the web harvesting seeds with which archivists may be familiar, such as those entered into Archive-It. They are similar in the sense that the user supplies a designated starting point from which a specific set of resources will be captured. And there is another similarity: the user can define parameters that should be applied to associated harvests.<sup>25</sup> The difference lies in the fact that SFM seeds do not target public web resources (i.e. web pages and their components). Rather, they aim to capture the output of an API: data that typically undergoes machine processing before being rendered in an interface.

Furthermore, there is another difference: SFM supports several different types of seeds, and each one of them behaves somewhat differently, in accord with the features afforded by the API that it is querying. It is important to note that the seed type is set at at the **collection** level in the user interface, and that in effect, the seed type is a collection attribute. This means that each collection will comprise a particular type of resource, one that format based. It is not possible, for instance, to combine tumblr blog, flickr posts, and tweets in the same collection. Nor can two different seed types be mixed in one collection; for example a twitter timeline and a twitter search seed cannot be joined side-by-side in the one collection.

---

<sup>24</sup> If the same seed (for example, a twitter timeline harvest for the seed @RealDonaldTrump) listed under two different collections, the database would record two different seed records, each with its own unique ID. It should also be noted that the twitter search harvest only allows harvesting from one seed at a time.

<sup>25</sup> For example, in Archive-It, the user can specify the harvester should capture just the seeded page, other pages on the same site, and/or external resources. Similarly, in SFM, the user can specify whether the harvester should capture just the structured data supplied by an API or related resources, such as embedded photos/video or linked webpages.

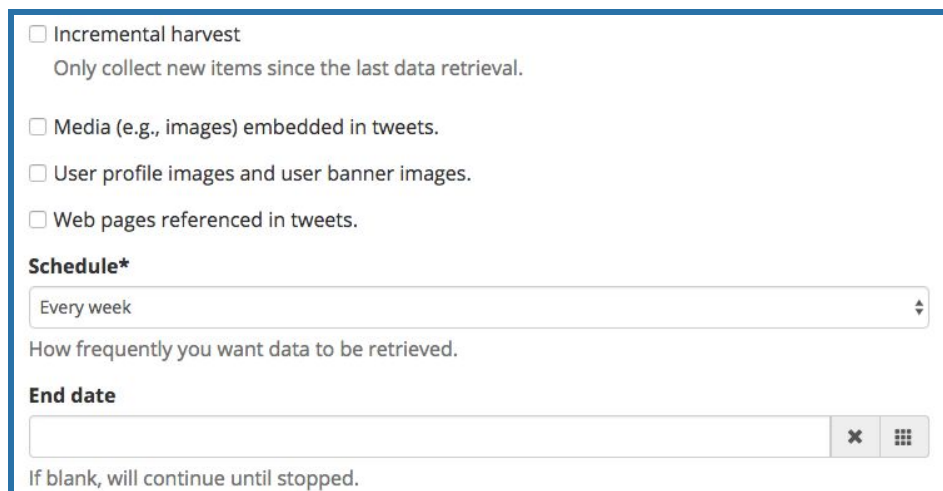
### *Practical Implications*

Archivists and other users benefit from understanding SFM's technical model because it shapes the ways in which they can collect, preserve, and access social media records. For example, the fact that each collection may only comprise seeds of a particular type imposes certain limits on collecting practice. On balance, these limits have more benefits than drawbacks. Not only do they contribute to the tool's overall ease of use, but they result in the gathering of consistent, orderly, and preservable digital objects. Nevertheless, they must be borne in mind by those using the tool, particularly in cases where there is some expectation that the data will be preserved for its continuing value, not just to serve an immediate research need.

For example, say an archivist would like to collect tweets, tumblr posts, and flickr gallery images from a person who has public accounts in all three services. Furthermore, assume the collecting organization would like to preserve these outside the system and to group them by their provenance, as the output of that person's activities and in relation to other digital materials generated by that person. In order to capture these three types of records (tweets, tumblr posts, and flickr images), the archivist will need to establish three separate SFM collections (a twitter timeline harvest, a tumblr blog post harvest, and a flickr user harvest). To group these SFM collections as a preservation object, staff would need to make a choice. They might group them in a single collection set, stating their common relationship by provenance in the note field of the collection set. Or, they might keep them in separate collection sets, then unite them externally when packaging them for deposit in a preservation system. In this case, staff members may need to export three sets of records, package them as a unit, then describe that package and its component parts, if they want to show their relationship to each other as the activities of a single individual.

In other words, decisions made at the collection set and collection level should be taken very carefully. An archives' staff members should carefully consider how they will shape their collections set and collections (By topic? By provenance? By some other factor?), basing decisions on long-term collecting objectives, descriptive practices, preservation capacities, and other factors they deem most relevant.

The decision about how to shape collections matters for another reason: Harvest parameters are set at the collection level (as shown in figure four) and are stored in the database as properties of the collection. In practical terms, this means that the harvesting of external objects and the frequency of harvesting are set on a collection basis, and that decisions made here will apply to each of the seeds that compose a collection.



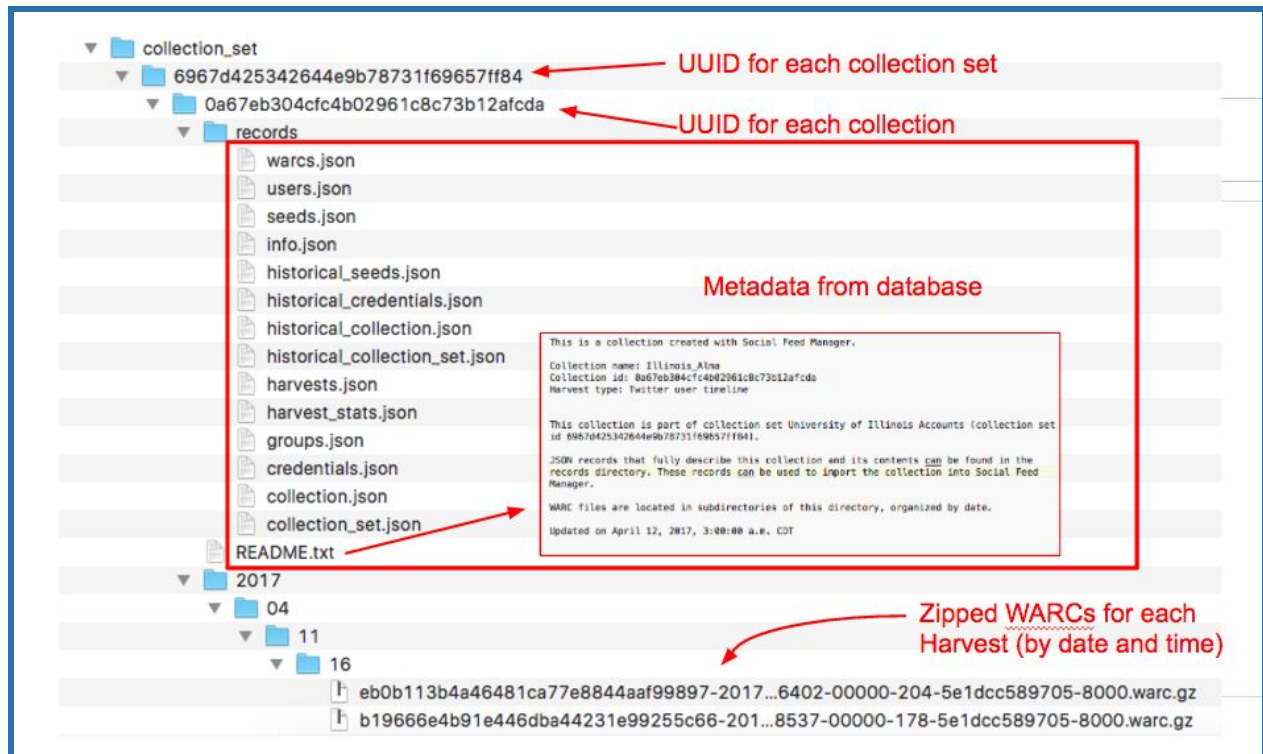
The screenshot shows a configuration panel with the following elements:

- Incremental harvest  
Only collect new items since the last data retrieval.
- Media (e.g., images) embedded in tweets.
- User profile images and user banner images.
- Web pages referenced in tweets.
- Schedule\***  
A dropdown menu currently set to "Every week".
- How frequently you want data to be retrieved.
- End date**  
An empty text input field with a clear button (x) and a help icon (grid).
- If blank, will continue until stopped.

**Figure Four: Harvest Options and Scheduling**

Looking back at figure three, each collection record may contain one or more harvests. In essence, a harvest is conducted whenever the SFM application is running and whenever a user or the system initiates a capture, based on the scheduling frequency noted on the collection record. When a harvest begins, SFM's messaging agent queues and issues request for the seeds that are currently active in that collection. It issues and API call to the social media service provider and copies the response to a WARC file, writing the precise JSON response as supplied by the API, along with the https headers. If the user has configured the collection to harvest web resources (such as embedded files and links to web pages), they are written to a separate WARC file. When harvesting has finished, the series of WARC's are written to a docker container or the file system (as specified in the `.env` file's configuration setting.)

It is critical to understand that for each harvest, the system captures both a warc file and some metadata about the warc. Each capture is stored in a very systematic way and with a JSON representation of the metadata that is stored in the database, as shown in figure five.



**Figure Five: SFM Data Storage on File System**

As additional harvests are completed, the application generates additional WARC files, storing them in a directory structure that reflects the harvest date. Metadata about the WARC's is stored in the database and, on a nightly basis, serialized into the JSON files, as shown in the screenshot above.

In effect, SFM functions like a 'minimalist repository,' where all relevant data and metadata is comprised in a single, semantically-transparent and modular package.<sup>26</sup> This helps ensure the portability of data between SFM instances; moving data for a collection set or collection to a new SFM instance is a simple matter of copying the containing folder and subfolders to a new location and running an import script. GW Libraries' decision to support this portability confers knock-on benefits, in that each collection or collection set is stored as a type of self-describing archival information packet. This is because SFM keeps an excellent record of its actions, as well as the actions of the users who configure the system right in the json files, fully serializing whatever information the database holds about a particular collection. To expand this point just a bit: SFM keeps records of configuration settings that were in place at a particular point in time. This data is kept in the historical tables, but is used to populate the json records contained in the records folder for each collection. For instance, if a collection contained ten seeds on May 10, 2016, but a user deleted three of those seeds and added six new ones on May 25, then added four more seeds on May 27, there would be three entries in the database: one for the current

<sup>26</sup> I would like to thank Justin Littman for suggesting this term and for sharing access to his unpublished draft explaining the concept.

settings (until they are replaced) and two in the analogous historical seeds table. All of this is mirrored as key/value pairs to the “historicals\_seeds” JSON file.<sup>27</sup> Similarly, if a user stops collecting or changes the description of a collection, that action is recorded in the database and serialized to the JSON file. Figure six provides a simple example of system change data stored as a ‘historical seed’ entry when a new seed was added to an existing collection.

The screenshot shows the Django administration interface for 'Change historical seed'. The form contains the following fields:

- History user:** sfmadmin
- History date:** Date: 2017-04-17, Time: 16:14:28
- History note:** Adding U of I Admissions Office Twitter to this Collection
- Seed id:** 8c963310888c4e2e83a789e8b88f53cd
- Collection:** <Collection 3 "Illinois\_Alma">
- Token:** uofiadmissions

A red box highlights the following JSON object:

```
{
  "fields": {
    "uid": "",
    "date_updated": "2017-04-17T21:14:28.636Z",
    "is_active": true,
    "collection": [
      "0a67eb304cfc4b02961c8c73b12afcd"
    ],
    "history_note": "Adding U of I Admissions Office Twitter to this Collection",
    "token": "uofiadmissions",
    "is_valid": true,
    "seed_id": "8c963310888c4e2e83a789e8b88f53cd",
    "date_added": "2017-04-17T21:14:28.631Z"
  },
  "model": "ui.seed"
}
```

**Figure Six: Configuration Setting in Change Log and in seeds.json File**

From an archivist’s perspective—as well as that of a future user of the records—this means that not only is SFM keeping a good record of provenance, it is making that record available in an easily preservable fashion, as an integral element of the package structure. The WARC and JSON files compose a submission information packet (SIP), to be used when the records are being ingested into another system. In essence, each collection set or collection contains many element of preservation metadata that one would need to preserve the bits over time and to allows users to make claims about their authenticity. For example, the JSON files include a record of prior collecting activities and an audit trail of actions taken by the people who have used the system to capture a particular set of social media posts. In fact, the packet could very nearly be considered to be ‘archival’—that is to say, an archival information packet (AIP). The only thing missing is additional descriptive metadata, which should be supplied by an archivist in an external descriptive system.

<sup>27</sup> SFM syncs database entries to JSON each night, or an administrator can sync on demand by running a serialization script from terminal.

While this point may seem rather academic, the fact that SFM keeps so much metadata about its actions allows us to differentiate between the social media posts that are captured and their value as records, in other words between a tweet and what might be called a ‘tweet record’.<sup>28</sup> Users can know that each tweet is a particular fixing of a social media post, an outcome from the interaction between decisions made by a human user and the technology capturing records, at a particular point in time.

This has important implications: First, it provides a wealth of structured information that makes the objects preservable in just the way they were captured. But even more so, it provides a basis for making future judgements about their authenticity or completeness of what has been captured, a topic discussed in more detail in the export options and packaging recommendations sections of this report.

While the data model that the SFM employs—tracking each action or harvest in a very hierarchical fashion—provides a well-structured data package, it also means that the data is stored in a complex folder structure, as shown in figure seven. (The implications of this for preservation and access are discussed more fully in the data exports and packaging sections of this report.)



**Figure Seven: Storage of WARCs by Harvest Date**

## Developing Collection Sets, Collections, Seeds, and Harvests

Before establishing collection sets, collections, seeds, and harvests, end users of the system will want to familiarize themselves with SFM’s user interface. The documentation’s User Guide provides an overview of the system features and describes the major considerations that must be taken into account.<sup>29</sup> It treats topics like establishing an account, adding API credentials, establishing collections sets and collections, and exporting data for processing and use in external tools or services. It also outlines the types of harvests that can be established and points to some important ethical and legal guidelines that should be respected, including

<sup>28</sup> I would like to thank David Dubin for suggesting this phrase, in the context of a discussion about a different project.

<sup>29</sup> See <http://sfm.readthedocs.io/en/latest/userguide.html#>.

restrictions on republishing datasets. Any user with reasonable technical facility should be able to navigate this guidance and begin harvesting records with no more than a few hours of time.

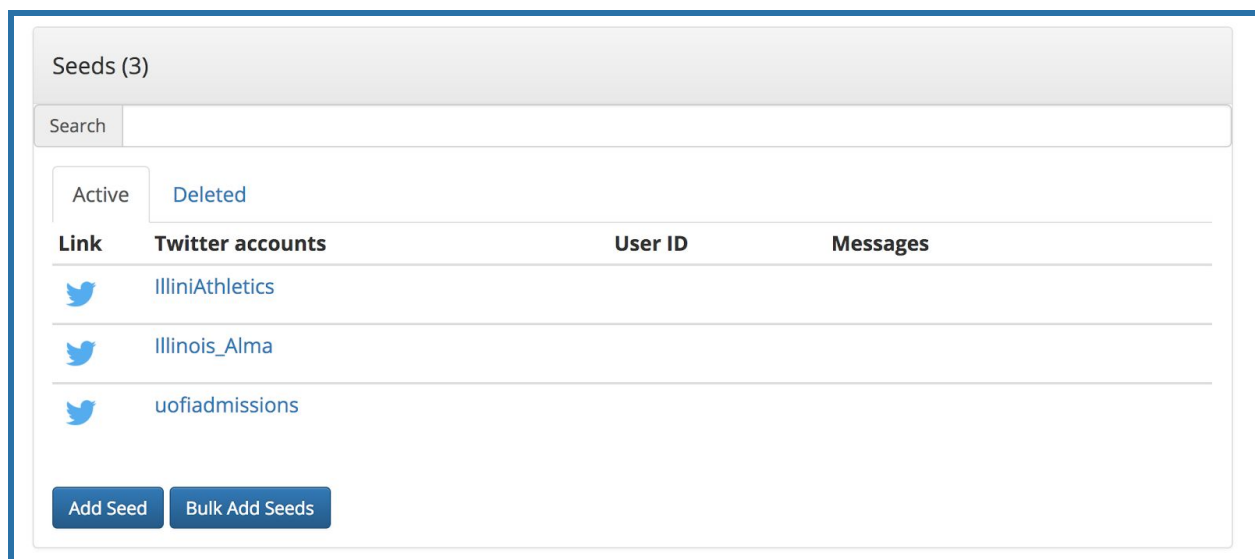
SFM currently supports the following collection/harvest types:

### *Twitter User Timeline*

This collection type collects tweets from a specific Twitter account or accounts. Several basic uses suggest themselves:

1. An archives could use this to collect social media sent from an account owned by the parent organization of the archives.
2. It might seek permission from a third party to record and store all of their tweets.
3. It could harvest of a third party's tweets without explicit consent, a strategy that may be most useful in the case of public figures or organizations whose tweets are of interest to a researcher or match the archives collections profile.

Twitter timeline collections allow the user to seed one or more user accounts into the interface, supporting several potential implementation strategies. For example, social media records from accounts that are related by topic or provenance could be grouped in a single collection; Figure eight shows some University level accounts grouped under a single timeline collection. All tweets from these three accounts will be grouped in one WARC file whenever the harvest runs.



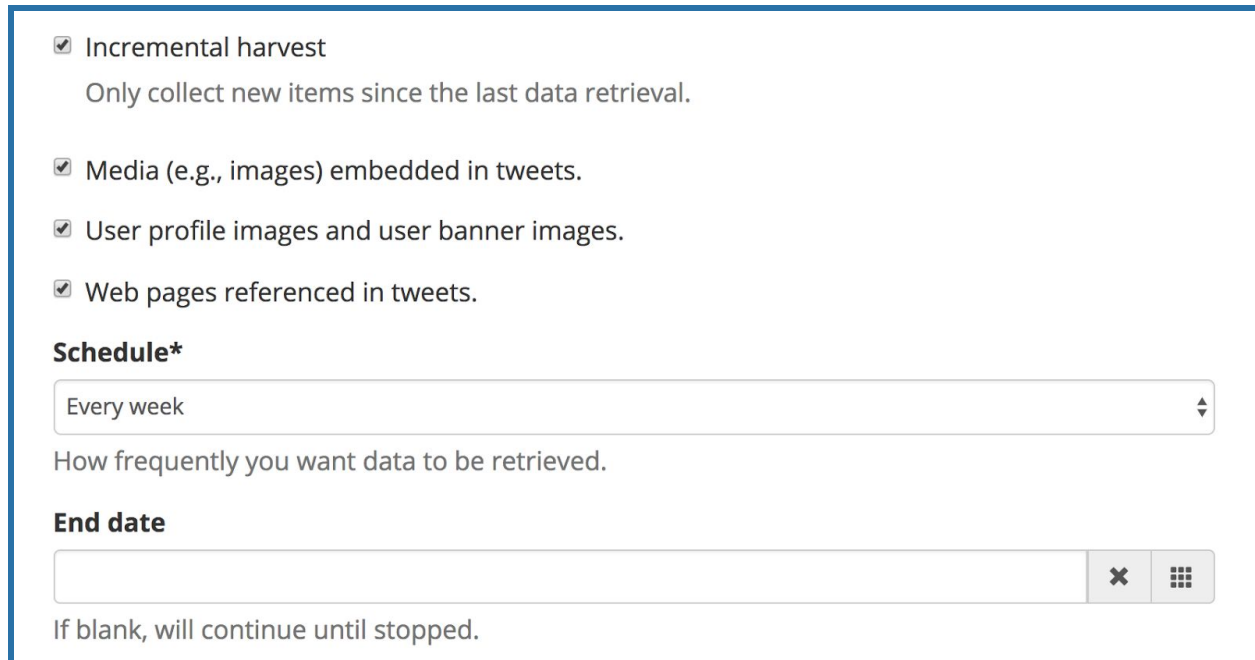
**Figure Eight: Twitter timeline harvest with multiple accounts.**

One caveat: SFM cannot harvest of all the tweets for an account that has tweeted more than 3,200 times, due to limits imposed by Twitter. To get a more complete record of such an account, the user would need to pursue some other methods, such as contacting the account owner and requesting a copy of the user's archive (which the account owner would need to export from Twitter). The archive would contain all of the tweet ids, so archivists could use



hydrator to capture the JSON representation of those tweets, if so desired.<sup>30</sup> After capturing the additional tweets, they would complement the SFM data. If the collecting organization wishes to preserve all of the tweets as a single dataset, they would need to be placed in an overall package structure and described properly in an external descriptive system, a topic discussed in more detail in the packaging section of this report. Ideally, a future version of SFM would allow import of tweets from a list of tweet ids (i.e. by incorporating hydrator or by referencing the twitter GET statuses/lookup endpoint.)

As another use case, collecting organizations could use timeline collections to target web resources for a known accounts that regularly includes embedded media or linked resources that are known to be of long-term value. In this case, it would make most sense to do incremental harvests, to avoid excessive duplication; the appropriate settings are shown in figure nine.



The image shows a configuration interface for a Twitter timeline harvest. It includes several checked options: 'Incremental harvest' (with a sub-note 'Only collect new items since the last data retrieval.'), 'Media (e.g., images) embedded in tweets.', 'User profile images and user banner images.', and 'Web pages referenced in tweets.'. Below these is a 'Schedule\*' dropdown menu set to 'Every week' with a sub-note 'How frequently you want data to be retrieved.'. At the bottom is an 'End date' input field, which is currently blank, with a sub-note 'If blank, will continue until stopped.'.

**Figure Nine: Twitter Timeline Harvest Configured for Incremental Harvests and Capture of Related Resources**

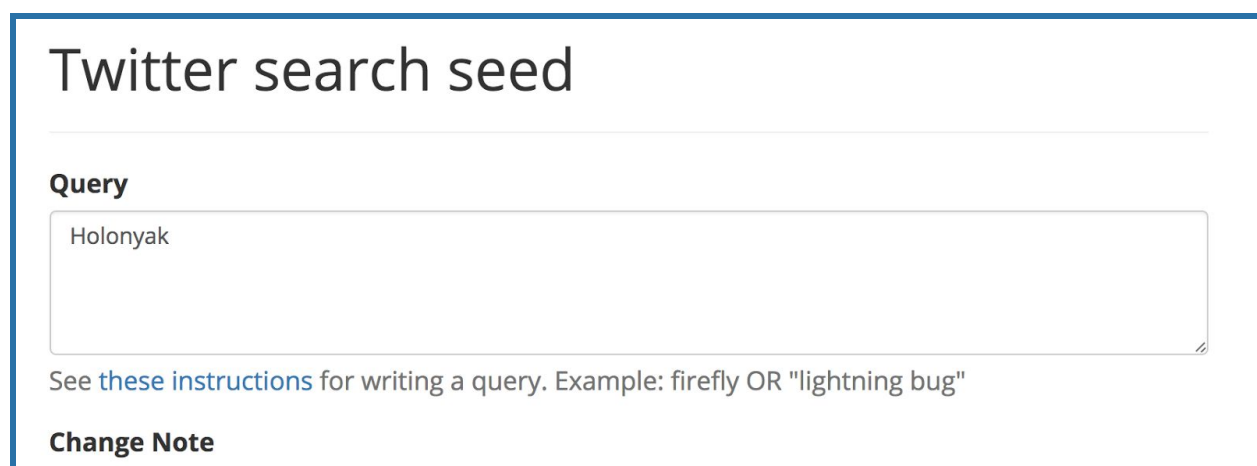
When harvesting web resources, data will accumulate rapidly and exponentially. For a sample account containing only 240 tweets, I ran both a tweet-only harvest and a harvest of tweets, web resources, profile pictures, and embedded media. The tweet-only harvest resulted in a compressed WARC that was only 78.7 KB in size; the harvest including the media included 4,459 web resources in compressed WARCs that were 102.3 MB in size--nearly 1,300 times

<sup>30</sup> Documenting the Now Project, *Hydrator Code Repository*, 2017. <https://github.com/DocNow/hydrator>.

larger than the JSON-only WARC. These factors must be borne in mind, especially for accounts that tweet more frequently than my sample account.

### *Twitter Search*

Twitter search harvests collect a selection of tweets from those posted in the last seven to nine days, matching the search terms specified. Unlike user timeline harvests, only one seed can be active at a time, and users should note that the results returned by the search API vary from those returned via the search box on Twitter. This collection type can be used to collect tweets concerning a term or hashtag known to be of interest to a researcher; for example, figure ten shows a simple search for tweets regarding a faculty member at the University of Illinois.



**Twitter search seed**

**Query**

Holonyak

See [these instructions](#) for writing a query. Example: firefly OR "lightning bug"

**Change Note**

**Figure Ten: Seeding a simple twitter search collection**

This search returned a very small set of data: 25 tweets. A search for a widely used hashtag will return a large amount of data; "#maga" (Make America Great Again) returned 17,778 tweets over two hours when it ran on May 3rd, 2017. The JSON data from this search composed over 14 MB of text, so a long-term harvest of common search terms could generate very large amounts of data, even if web resources and embeds are not included. If web resources were included and a proportional amount of data were generated for this harvest as in the 240 tweet harvest mentioned above, over 18 Gigabytes of data would have been captured.

Twitter search collections can be used as an effective way to document emerging events or memes, when the importance of the event or meme becomes known after the fact. Consider that an important local event or controversy takes place, and a wide range of people and organizations begin tweeting about it. However, the archives staff members do not know a relevant search term until a day or two after the fact. They could set up a twitter search collection in SFM, reaching back up to seven or nine days into the record of tweets matching the search term.

This backwards glance differentiates search harvests from the filter harvests discussed below, which will include ONLY materials from the live Twitter stream. But again, users should be aware that the data supplied via the API will not represent all of the relevant tweets, only those supplied by Twitter's undocumented algorithm.

With twitter search harvests, considerable care should be placed on how the search term is defined. As noted in SFM documentation and on the Twitter website, simple Boolean-style queries are allowed. The supported syntax is a bit limited since grouping with parentheses is not supported. A query for `#cumarchforscience OR @ScienceMarchCU OR from:ScienceMarchCU` would include tweets that either (1) include the hashtag, (2) mention the account, or (3) were sent from the account. A query for `#cumarchforscience AND from:ScienceMarchCU` would only include tweets that match both parameters. The more carefully a user or archivist constructs the query, the more likely that they will capture meaningful response data. As with other harvest types, web resources can be captured, if desired.

### *Twitter Sample*

This collection type captures a Twitter-defined random sample of all tweets in real time, somewhere from .5 to 1% of all tweets being posted. By grabbing a stream of all tweets, an end user may be able to analyze the entire set of topics being discussed on twitter at any given time, particularly if they have the technical facility to analyze the data with natural language processing, entity extraction, or topic modeling software. Archives should carefully consider whether such an amorphous assemblage belongs in their preservation repository. It seems doubtful, since the dataset will lack any topical focus. But it might have some value as a point in time snapshot following events of world-historical prominence, particularly those relate to a repository's topical strengths.

### *Twitter Filter*

As noted above, collections that use the Twitter filter API differ from those collected via the search API, since tweets are captured as they are posted real time, not from the recent past. But there is a less obvious difference. Users supply a different set of seed criteria to determine the exact nature of the data that is returned, as shown in figure eleven, than they do when seeding Twitter search collections.

## Add Twitter filter seed

**Track**

Separate keywords and phrases with commas. See Twitter [track](#) for more information.

**Follow**

Use commas to separate user IDs (e.g. 1233718,6378678) of accounts whose tweets, retweets, and replies will be collected. See Twitter [follow](#) documentation for a full list of what is returned.

**Locations**

**Figure Eleven: Defining twitter filter seed to collect tweets about or by James Fallows.**

In the “track field”, commas function as a logical OR operators, and each word is treated as a discrete and bounded search term. Stemming is not automatic, so the simple filter `base` will not match tweets with the term ‘baseball’ but only those including the discrete word ‘base’. For this reason, users should provide all expected variations of a term that they wish to track. Users should also recognize that any spaces between the terms will be treated by the API as a logical AND operator. So, a filter specified as `bass fish` will only match tweets that contain both of those terms, but not one that contains the terms ‘bass’ and ‘guitar’ unless a user tweets something like this: “I just caught a huge fish while playing my bass guitar.” Given the API’s complexities, users should follow a trial and error approach, confirming that the syntax works in a small, controlled harvest, before committing to a large, long-term one likely to create a massive dataset.

There are a few others wrinkles to keep in mind when seeding twitter filter collections. The term must be included in the body of the tweet; it will not match screen names or account names. So a search for `jamesfallows` will not necessarily capture tweets sent from that account, unless the follow attribute is set to track the ID of a particular account. . Second, the track, follow, and location criteria are also treated as a logical OR operations. In figure eleven, tweets that include the text “jamesfallows” or “fallows”, are captured by the track parameters, and those posted by by the account with the specified user id (which is associated with the screen name jamesfallows), are captured by the follow attribute.

### *Flickr User*

This collection type will capture posts and metadata from specific Flickr accounts. Optionally, copies of photographs can also be harvested, at the resolution or resolutions specified on the collection details screen. Once a Flickr user collection has been established, one or more user ids need to be seeded, typically after looking up the user ID for the photos associated with a particular user, as shown in figure twelve.

**Add Flickr user seed**

**Username**

A string name for the user account. Finding this on the Flickr website can be confusing, so see NSID below.

**NSID**

An unchanging identifier for a user account, e.g., 80136838@N05. To find the NSID for a user account, use idGettr.

**Change Note**

Further information about this addition.

**idGettr**

Use the URL of your photostream to find the Flickr ID number (also works for groups).

id: 45214447@N07

**Figure Twelve: Adding a seed for a Flickr user collection**

When establishing collections, seeds, and harvests, archives staff will need to carefully define policies regarding the sharing of metadata and photographs harvested from Flickr, in line with the rules specified by the API terms of use.<sup>31</sup> In particular, SFM users and repositories must note that the owners of the photographs retain all rights to the images, and that they must comply with the rights requirements pertaining to each image that is hosted in the service. In the JSON response that the API returns, a license code is provided as a simple integer, corresponding to the license options that Flickr allows its users to choose, as shown in figure thirteen.

<sup>31</sup> <https://www.flickr.com/services/api/tos/>.

```

<licenses>
<license id="0" name="All Rights Reserved" url="" />
<license id="1" name="Attribution-NonCommercial-ShareAlike License" url="http://creativecommons.org/licenses/by-nc-sa/2.0/" />
<license id="2" name="Attribution-NonCommercial License" url="http://creativecommons.org/licenses/by-nc/2.0/" />
<license id="3" name="Attribution-NonCommercial-NoDerivs License" url="http://creativecommons.org/licenses/by-nc-nd/2.0/" />
<license id="4" name="Attribution License" url="http://creativecommons.org/licenses/by/2.0/" />
<license id="5" name="Attribution-ShareAlike License" url="http://creativecommons.org/licenses/by-sa/2.0/" />
<license id="6" name="Attribution-NoDerivs License" url="http://creativecommons.org/licenses/by-nd/2.0/" />
<license id="7" name="No known copyright restrictions" url="http://flickr.com/commons/usage/" />
<license id="8" name="United States Government Work" url="http://www.usa.gov/copyright.shtml" />
</licenses>

```

**Figure Thirteen: Flickr license options.**

Source: <https://www.flickr.com/services/api/flickr.photos.licenses.getInfo.html>

In addition to complying with the owner's wishes, SFM users must meet the terms of service for Flickr's API. In practical terms, collecting organizations may wish to provide users with a set of photo ID's. This would allow users to retrieve photos from Flickr, which the collecting organizations keep preservation copies of the metadata and images.

The archives should proceed with caution when researchers harvest photographs for a personal research project. If the photographs are owned by the person or organization donating them, the archives would be on solid ground in accepting the collection; ideally ownership would be transferred or use would be licensed under a deed of gift or other legal instrument, as with any other digital photograph or asset. If, on the other hand, the photographs are owned by a third party and were harvested by another person using SFM for personal research, that user would not have the right to donate the copies to the archives, so the archives could only accept, preserve, or provide access to them after undertaking a risk assessment along the lines suggested by Aprille McKay.<sup>32</sup> If the photos are provided under a creative commons license, that would certainly facilitate downstream reuse of the preservation copies.

### *Weibo Timeline and Weibo Search*

While I was unable to test these collection types, given my lack of knowledge in the Chinese language, they operate in a somewhat different way than Twitter collections. Yecheng Tan has provided a useful guide in a blog post on the SFM website.<sup>33</sup>

Weibo can be translated as 'microblogging,' and the service is sometimes portrayed as a cross between Twitter and Facebook. Users compose short posts and share them with friends. The Weibo timeline API allows users to (possibly) collect posts from their friends' accounts as well as their own. If you follow someone, you might get some of their weibos (posts) in your timeline.

This suggests a potential collecting and documentation strategy: an archivist or end user might directly friend Weibo users whom they wish to document, then collect the timeline for that

<sup>32</sup> McKay, "Managing Rights and Permissions," 176-210.

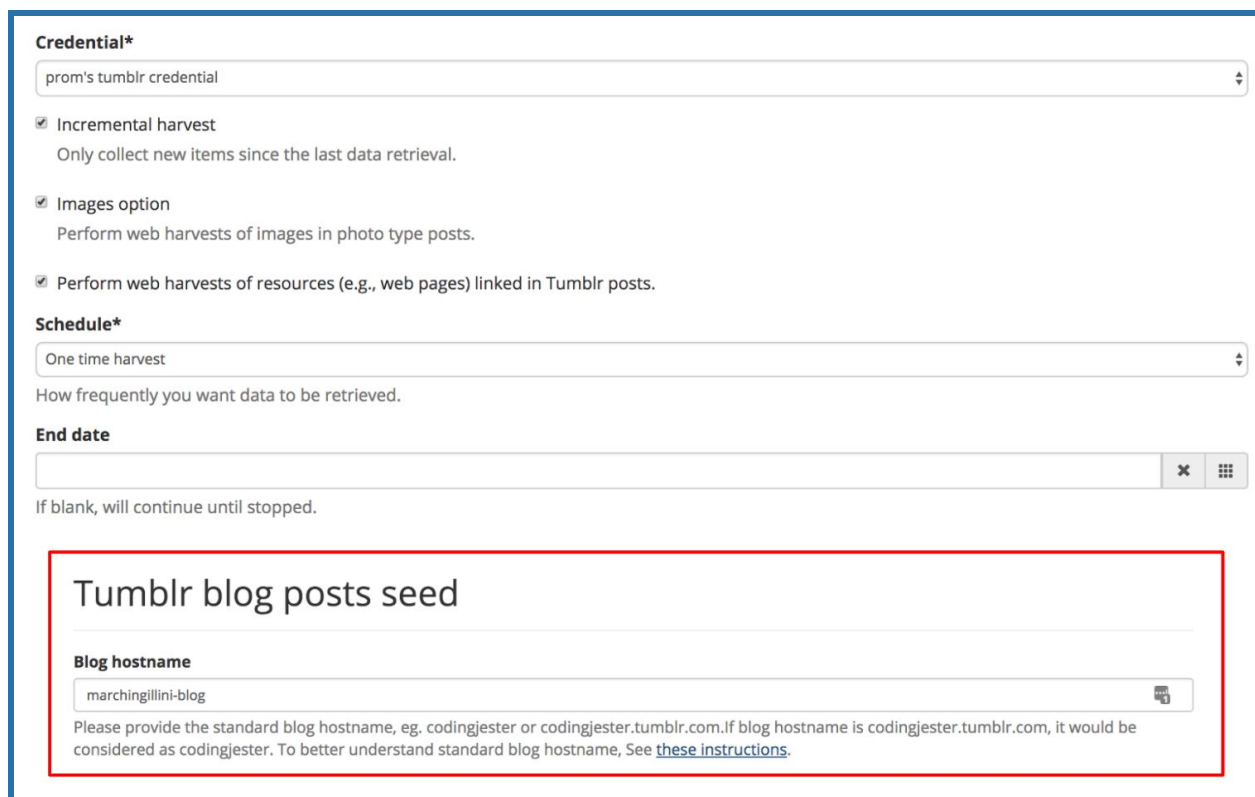
<sup>33</sup> Tan, "Weibo API Guide," <https://gwu-libraries.github.io/sfm-ui/posts/2016-04-26-weibo-api-guide>.

account. By seeking permission from the targeted users, the archives would be very transparent regarding its collecting actions, perhaps building trust with a particular community that seeks to document its own activities. If a future version of Facebook's API allows for easier data collection, perhaps such a strategy could be likewise applied.

With Weibo, such a strategy would raise significant ethical questions, which should be considered fully in light of the people and organizations whose materials are being collected, as well as the social and political context of their activities; SFM's project's Collection Development Guidelines, which suggest a set of questions that repositories should consider when launching collecting projects.<sup>34</sup>

### *Tumblr Blog Posts*

This collection type will capture posts from a defined user account in Tumblr. As with the Twitter user timeline collections, harvest can be incremental or include embedded media, if desired. Figure fourteen shows a user setting collection details and entering a tumblr collection seed.



The screenshot displays a web form for configuring a Tumblr collection. The form is organized into several sections:

- Credential\***: A dropdown menu with the selected value "prom's tumblr credential".
- Incremental harvest**: A checked checkbox with the subtext "Only collect new items since the last data retrieval."
- Images option**: A checked checkbox with the subtext "Perform web harvests of images in photo type posts."
- Perform web harvests of resources**: A checked checkbox with the subtext "(e.g., web pages) linked in Tumblr posts."
- Schedule\***: A dropdown menu with the selected value "One time harvest". Below it is the subtext "How frequently you want data to be retrieved."
- End date**: An empty date input field with a clear (x) and calendar icon button. Below it is the subtext "If blank, will continue until stopped."
- Tumblr blog posts seed**: A large text input field containing the text "Tumblr blog posts seed".
- Blog hostname**: A text input field containing "marchingillini-blog" and a help icon. Below it is the subtext: "Please provide the standard blog hostname, eg. codingjester or codingjester.tumblr.com. If blog hostname is codingjester.tumblr.com, it would be considered as codingjester. To better understand standard blog hostname, See [these instructions](#)."

**Figure Fourteen: Tumblr Collection details and seed settings**

<sup>34</sup> <https://gwu-libraries.github.io/sfm-ui/resources/guidelines>.

## Web Resources

As noted above, nearly all of the harvest types allow SFM users to collect secondary materials that are linked from or embedded in social media posts, in which case Heretrix captures and the harvested resources and writes them to a WARC file. This is written directly into the collection directory on disk, alongside the JSON data returned by the social media service's API, but in a separate WARC, as illustrated in figure fifteen.

The screenshot displays the Social Feed Manager interface. At the top, there is a 'Seeds (1)' section with a search bar and filter buttons for 'Active' and 'Deleted'. Below this, a 'Link' section shows 'Twitter accounts' with a search term 'chrisprom' and buttons for 'Add Seed' and 'Bulk Add Seeds'. The main section is 'Harvests (2)', which contains a table with columns: Type, Requested, Updated/Completed, Status, Stats, and Messages. Two harvests are listed:

Type	Requested	Updated/Completed	Status	Stats	Messages
Web	May 3, 2017, 12:00:24 a.m. CDT	May 3, 2017, 12:30:39 a.m. CDT	Success	4,459 web resources	
Twitter user timeline	May 3, 2017, 12:00:15 a.m. CDT	May 3, 2017, 12:00:23 a.m. CDT	Success	240 tweets	

Two harvest details are shown in red boxes:

- Twitter Harvest:**
  - Id:** d41112d88f8246a59c7b3744756e9b8b
  - Requested:** May 3, 2017, 12:00:15 a.m. CDT
  - Started:** May 3, 2017, 12:00:15 a.m. CDT
  - Ended:** May 3, 2017, 12:00:23 a.m. CDT
  - Updated:** May 3, 2017, 12:00:27 a.m. CDT
  - Status:** Success
  - Performed by:** Twitter Harvester on 54cf6c6b401 (14)
  - Harvest type:** twitter\_user\_timeline
  - Stats:**
    - tweets: 240
  - WARCs:** 1 file (73.4 KB)
  - This harvest requested:**
    - Web harvest (May 3, 2017, 12:00:24 a.m. CDT)
- Web Harvest:**
  - Id:** e995158e49a348d5bb81e0bac30c7ffa
  - Requested:** May 3, 2017, 12:00:24 a.m. CDT
  - Started:** May 3, 2017, 12:00:23 a.m. CDT
  - Ended:** May 3, 2017, 12:30:39 a.m. CDT
  - Updated:** May 3, 2017, 12:30:40 a.m. CDT
  - Status:** Success
  - Performed by:** Web Harvester on 687f9f25e5aa (14)
  - Harvest type:** web
  - Stats:**
    - web resources: 4,459
  - WARCs:** 1 file (102.3 MB)
  - This harvest requested by:** Twitter user timeline harvest (May 3, 2017, 12:00:15 a.m. CDT)

**Figure Fifteen: Web and Twitter timeline harvest data collected for one-time capture**

The ability to collect of web materials that are related to tweets opens up interesting collection opportunities. One can envision (through a glass, and darkly) the multidimensional record space mentioned earlier in this report. Certainly, tweets or other social media posts will make much more sense on an individual basis, if users can access copies of referenced or embedded object. However, archival repositories and users will want to be very careful in their use of the web harvest options. Even in the case of small web harvests, the amount of data that is collected can rapidly grow to very large proportions, particularly in cases where a large number of social media records are being captured. Current version of SFM cannot easily complete large scale web harvests. More testing and development would need to be done, if this experimental feature were to be expanded.

Collecting web resources is, in any case, a secondary objective with SFM. Heretrix was added into the project as useful supplement to the work formally supported by NHPRC, which focused around API-based collecting and the development of datasets. In other words, this work would



need future attention in a future grant cycle, were it to be considered important to the community.

In the meantime, users should note the captured web resources will require considerable preservation and access planning. For now, there are many open questions. While it is easy to preserve the WARC bits, links between individual social media records and referenced web resources are implicit and buried in a bitstream; how should they be exposed to users or leveraged for access? In the case of long-term or repeated harvests, many WARC files will be generated; how should they be packaged for dissemination. And what about rendering? While there are easy to implement desktop tools for displaying the preserved WARC files (such as Webrecorder Player and WAIL), it is unclear how or whether these tools could be used to display web resources captured by SFM.<sup>35</sup> Additional testing is necessary, and the future of those tools is a bit uncertain as they are tied to specific projects and developers. SFM's documentation provides good instructions for using JSON processing tools built into SFM,<sup>36</sup> but currently SFM does not include many tools for accessing or using WARC files created by Heretrix.

Speaking personally, I feel that the historical record would be greatly enriched were it possible to SFM or SFM like methods to systematically capture, preserve, discover and access web resources alongside the social media records that reference them. Other approaches to web archiving, with tools such Archive-It, make it relatively easy to collect a certain site or set of sites, starting from a seed URL. But given the prominence of social media, and the fact that materials relating to multiple topics live in a complex web of relationships and social environment, one can certainly envision a case for using SFM or a tool like SFM to capture web resources that relate to a common topic or activity, yet are dispersed across dozens, hundreds, or even thousands of domains. This transformative use may be one of the most salutary if unexpected long term benefits of the SFM project work supported by NHPRC, even if it was not part of the formal project objectives.

## Shaping Local Services

Once archival staff understand SFM's basic functions, they can make decisions about how the tool will be used. Some considerations are worth extra attention by repository staff before they or other users begin establishing collection sets, collections, and seeds, particularly if the records will be saved for future research value in an archival or special collections library.

Users should note, for instance, that SFM is structured with a relatively flat data model, whereby a collection belongs to only one collection set. It is not possible to define subcollections or otherwise differentiate or group seeds within a collection; nor can a collection be shared across

---

<sup>35</sup> Ilya Kreymer, *Webrecorder Player Code Repository*, 2017, <https://github.com/webrecorder/webrecorderplayer-electron>; Mat Kelly, "Web Archiving Integration Layer (WAIL)," <https://machawk1.github.io/wail/>.

<sup>36</sup> See <https://sfm.readthedocs.io/en/latest/processing.html>.

two collection sets. For an archivist who may be accustomed to classifying materials with increasing granularity, or in providing cross-cutting access points, these limits should be borne in mind.

After all, SFM's main purpose is to capture records, not to describe them. Nor is it intended to help repositories develop a classification structure that relates collections to each other or to other records that the same person or organization might have created. Both of these tasks should be completed in external systems, such as a repository's collection management software, not inside SFM. Yet the collecting organization will reap long-term dividends by using the descriptive fields that it makes available. Users should pay particular attention to the ways in which they organize, name, and describe collection sets and collections. Since people with access to SFM's user interface can generate collection exports (but not collection set exports), archivists will want to establish collections with an eye to how they will be described and preserved.

For instance, an archives might establish a collection set to harvest the timelines of accounts managed by the institution, then group seeds in collections thereunder, to reflect a shared provenance. Materials harvested under that collection set could be described and packaged and described in aggregate reflecting their provenance. The exported social media data for a collection would compose a single archival information packet, one that could be uploaded to a local digital preservation repository.<sup>37</sup>

Consider the example of an archival repository that has previously arranged the files of a campus unit or department, say the College of Liberal Arts and Sciences, within a single collection description, with links to particular record series that were generated by the College. The archives has previously developed a single authority record describing the College, and has linked that authority record to series level records descriptions (such as the Dean's Subject File, Executive Committee Minutes, Press Releases, Personnel Files; or the Subject Files of various departments found in the College). Social media records are simply another record of the College's activities. The archives could then group social media feeds for individual units in the college as a single collection, adding a seed for each twitter account managed by the college or one of its subunits. This would aggregate similar records together and would allow them to be exported as a single archival information packet, one that could then be described in an archival management system just like any other record of the college's activities. Figure sixteen shows a sample descriptive record for a packet of such records, as collected by SFM.

---

<sup>37</sup> SFM version 1.8 and higher support moving collections to other collection sets, so related collections could be grouped after the fact into a single collection set, then exported for external preservation, if a different organizational method is chosen by the repository.

### College of Liberal Arts and Sciences Twitter Timelines, 2017- | University of Illinois Archives

**Title:** College of Liberal Arts and Sciences Twitter Timelines, 2017-

**Series Number:** 15/1/840

**Volume:** 12.0 megabytes

**Arrangement**  
Records are kept chronologically by account, as produced by the Social Feed Manager, in their original JSON format.

**Creator(s)**  
[University of Illinois at Urbana-Champaign. College of Liberal Arts and Sciences](#)

**Administrative History of Creating Unit**

**Access Restrictions**  
Tweet ID datasets are made available via our Digital Library. original JSON data or spreadsheets are available by contacting archives, and may be used under the terms of the twitter. See <http://twitter.com/rules>.

[Email us about these publications](#) | [Print this information](#)

**Description:** Captured tweets from accounts managed by the College of Liberal Arts and Sciences or its component units, including the history department and the writers workshop.

**Access Restriction:** Tweet ID datasets are made available via our Digital Library. . . . [more](#)

#### College of LAS Twitter Timelines

**Collection name\***

**Description**  
Twitter timelines from accounts managed by the College of Liberal Arts and Sciences, or its component units

Seeds (7)			
Active	Deleted		
Link	Twitter accounts	User ID	Messages
	Chemistry@UC		
	History@LAS		
	@BioIllinois		
	IG@Illinois		
	LAS@Illinois		
	LAS@Academics		
	workshop@UC		

**Figure Sixteen: Sample Collection Record and SFM Record for Archival Packet**

The same archives may wish to develop a different implementation model for capturing social media posts by faculty who have agreed to donate their personal timelines to the Archives. In this case, a single collection could be established for each faculty member, in a collection set that is dedicated to Faculty Social Media. That would allow for the export of social media posts by a particular faculty member at some future point in time, in which case they could be integrated into a collection description for that faculty member's personal and professional papers, including other records such as subject files, correspondence (including email) or whatever other series reflect that faculty member's activities.

Similar arrangements could be made in the case of student organizations: a parent collection set for social media records from those groups, and collections grouping twitter timelines or search captures related to particular student groups or campus issues.

Yet other options suggest themselves. Since SFM supports the capture of records that support faculty or student research projects, all such projects could be segregated from the collections being generated and gathered (by archives staff) as an institutional record. The faculty and student captures could then be kept only for personal use under research projects, or, if the faculty and students agree, accessioned into the archives and managed under terms that respect the legal and ethical requirements that may pertain.

## Use cases

The end user documentation for SFM provides good instructions for using the software to establish collection sets and for shaping the nature of the collections that are included within them. It is a relatively simple matter to add new collections sets, collections, and seeds and to set the parameters or scope for a particular collecting project, whether limited term or ongoing.

However, those using the system should be aware that the specific decisions they make in structuring the collection sets, users, and seeds will shape the collections in a particular fashion, affecting their ability to preserve, describe, and provide access to them in the long term. While some of these issues were hinted at above, this section of the report provides some more targeted descriptions of particular ways that the repositories might wish to use SFM.

### GW Libraries User Stories

First, archivists and collection managers should thoroughly review the user stories developed by the GW Libraries project staff, which describe five basic scenarios.<sup>38</sup> Each of these describes a series of steps that a particular person would follow in using the tool to accomplish a defined objective. They are specific to GW Libraries and developed early in the project. While they inform the ways in which GW Libraries shaped the tool during the project, they continue to shed light on some questions that all repositories should ask when considering how to implement and support SFM in their own institutions. The GW Libraries user stories describe some of the specific tasks that repositories and users will want or need to complete using SFM.

- *Basic researcher*: An undergraduate researcher using SFM to gather sources for a course paper, with no expectation that the harvested data will be retained, browsing the data in SFM itself without exporting information for further analysis or expecting that it will be preserved by the archives or Library.
- *Event capture*: A faculty member seeks to capture records concerning the response of non-governmental organizations to a natural disaster, using them
- *Archivist*: An archivist collects Twitter, Flickr, and Tumblr posts under agreement with an organization whose records have been donated to the archives.
- *Future researchers*: Describes a student and faculty member who use a prospective search interface to query SFM for relevant social media collections, then export them and mine them using other software.
- *Collection review*: Describes staff from a collecting organization auditing the use of SFM and making decisions to transfer collection ownership, preserve materials outside SFM, or otherwise administer data collected by SFM.

---

<sup>38</sup> <https://gwu-libraries.github.io/sfm-ui/about/user-stories>.

Thinking more broadly about its potential uses, and knowing the some elements of the GW Libraries user stories were subsequently baked into the tool's current feature set, I see three general ways in which repositories might implement SFM:

1. Using it for records management purposes
2. Installing it for use by archival staff to proactively build collections for future research value.
3. Managing SFM as part of a research consultation service within the Library; and .

While these three approaches can complement each other and certainly are not mutually exclusive, they should be implemented in an appropriate management framework and with staffing or other support services, to ensure that the tool is made more effective for its intended uses. Accordingly, each of the following two subsections describes a general implementation scenario and highlights some management factors to consider should an institution wish to use it for the described purposes.

### **Institution-Led Collecting**

As noted in SFM's Collection Development Guidelines, archivists and curators may wish to use Social Feed Manager to collect social media collections that become part of their long-term holdings.<sup>39</sup> While this was not the original purpose of the tool in its early versions, and while it raises potential ethical and legal issues that will need to be managed in a policy framework, this section of the report outlines a few potential institution-led collecting efforts that repositories may wish to consider.

#### *Facilitating Records and Information Compliance*

SFM could be used as a cost effective means to copy and store social media posts sent by state agencies, public universities, or other organizations that have a need to ensure compliance with legal requirements or records management imperatives. By configuring one or more account timeline harvests, and organization would keep a potentially-complete record of all activity undertaken by the account. The institution would need to define policies for the retention and long term disposition of the materials, perhaps in conjunction with an archivist or member of legal counsel, and it is likely that some of the records, if the not the entire stream of activity, might have long term administrative or research value to the institution.

If an institution is considering the use of using SFM for records management purposes, staff members should bear in mind that SFM may not capture everything issued by a particular account or related to a particular search. It will certainly capture most traffic, but 100% capture is not guaranteed. This is less a limitation of SFM than of the free APIs that SFM uses. They simply do not guarantee complete coverage. If the collecting organization wishes to ensure the

---

<sup>39</sup> George Washington University Library, "Building Social Media Archives: Collection Development Guidelines," <https://gwu-libraries.github.io/sfm-ui/resources/guidelines>.

highest levels of compliance and capture, staff will need to contract with an external service, such as ArchiveSocial or DiscoverText. Nevertheless, SFM may have a records management role to play in cases where records are being captured less for legal reasons than historical or administrative ones, but where some compliance needs also exists.

### *Capturing Institution-Related Materials*

Institutional archives, such as a government, university and business archives, bear a responsibility to preserve records generated by their parent organization. While social media collecting is likely to be a new area for most repositories, SFM's twitter timeline harvest feature could be used to establish a one time or incrementally-updated feed of new tweets from one of more user accounts that are managed by the parent organization. The following possibilities suggest themselves:

- Using one collection set to collect an organization's records: In the case of smaller organizations with relatively few social media records, the organization might establish a collection set with a title like "[Organization Name] Social Media Records". Within this collection set, three collections could be established, one for twitter user timelines, one for flickr posts, and one for tumblr blogs. Within each collection, all streams across the organization could be listed, as shown in figure sixteen.

The screenshot displays the 'University-wide Twitter Account Timelines' collection in the Social Feed Manager. The main interface shows the collection is active, with a 'Turn off' button and 'Edit' and 'Export' options. A 'Next harvest' date is set for May 3, 2017, at 5:10:32 p.m. CDT. The 'Stats' section indicates 1 file (4.2 MB) of data collected, with 9,693 tweets. A 'Details' link is highlighted with a red arrow, pointing to a modal window that shows the collection's name and a description: 'This collection includes all twitter accounts that are managed by units of the University of Illinois at Urbana-Champaign.' Below the modal, there is a table of seeds and a table of harvests.

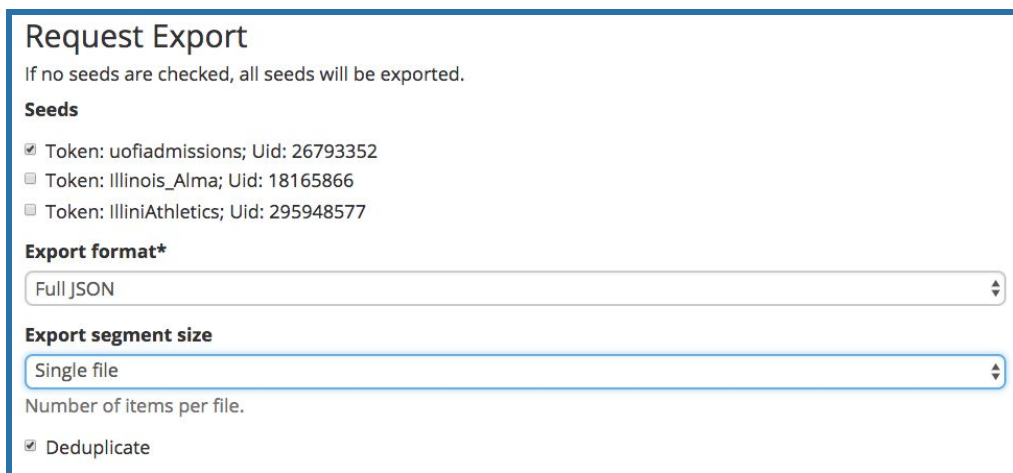
Link	Twitter accounts	User ID
	IlliniAthletics	295948577
	Illinois_Alma	18165866
	uofiadmissions	26793352

Type	Requested	Updated/Completed	Status	Stats	Messages
Twitter user timeline	May 3, 2017, 1:16:07 a.m. CDT	May 3, 2017, 1:16:59 a.m. CDT	Success	9,693 tweets	

**Figure Sixteen: Collection Aggregating Twitter Accounts that Share a Common Provenance**

The advantage to this approach lies in the fact it is relatively low-effort and low-maintenance: As new accounts are identified, they are simply added to the seed lists. And (assuming none of the account tweets more than 3,000 times per month), harvests could be run infrequently, resulting in a relatively low number of warcs to be preserved. However, a few potential drawbacks obtain. Notably, data from multiple twitter accounts will be included in each WARC. When staff members export data from SFM, they should consider how to integrate the exported data into a classification system and repository infrastructure. As a low-effort approach, the entire collection set could be packaged and preserved as a discrete digital object and described as a collection in a collection management system. If this is done, it will be much more difficult to provide some kind of series breakdown or to separate records by provenance, unless staff members create separate exports. If, on the other hand, an archives wants to separately classify records by accounts, it could filter exports by seed, as demonstrated in figure seventeen. This method would only be suitable in cases where webresources have not been harvested, since linked and embedded resources are not currently included in the exported dataset. Regardless, the “preservation by export” method would sacrifice some of the metadata that SFM keeps regarding its actions, as discussed in more detail below.



**Request Export**

If no seeds are checked, all seeds will be exported.

**Seeds**

- Token: uofiadmissions; Uid: 26793352
- Token: Illinois\_Alma; Uid: 18165866
- Token: IlliniAthletics; Uid: 295948577

**Export format\***

Full JSON

**Export segment size**

Single file

Number of items per file.

Deduplicate

**Figure Seventeen: Exporting data for a defined account from a collection**

- *Organization by records creator:* As a counterpoint, an archives may wish to establish separate collections for each account or for each records creating entity within an organization. Obviously, this would require more upfront effort, but collections could be described with more granularity within SFM, before being copied into the preservation repository and described in an archival management system. As new materials are captured in the system, the new harvests could be added as accruals, within the parent folder. This approach also holds a few potential downsides. In particular, collecting organizations should note that each collection will hold objects of only one type. If a unit uses all three channels (Twitter, Tumblr, and Flickr), and an archives wants to organize

records sharing this provenance, they would need to undertake some external packaging and addition description, before ingesting the files into a preservation repository.

- *Function-driven collecting*: This is a hybrid approach, combining some of the features of the others just described. If an institution uses a documentation strategies approach, accounts relating to public relations could be grouped together, accounts from academic departments in another, and institutionally affiliated student organization in a third.<sup>40</sup> The provisos mentioned above should be borne in mind, since the specific ways collection sets and collections are established will affect export, packaging, description, and preservation potentials that might be afforded to them outside of SFM.

### *Documenting Events*

Many repositories may wish to harvest materials documenting particular events or controversies, particular if they expect that their future user community will have an interest in the event. While the particular events being documented might range from the mundane (a graduation ceremony, perhaps) to the controversial (a violent act or protest), SFM's twitter search and filter harvests provide tools that archivists can use to capture data for future research and preservation.

As a relatively simple example, the University of Illinois Archives collected records for a Science March, using the twitter search harvester. We knew about relevant accounts and hashtags ahead of time, and by carefully constructing a search string, we gathered many records including some tweets and web resources that had been posted prior to the event.

This may seem like a subtle point, but the 'recordness' of this collection lies in the University of Illinois Archives' choice to collect the records, not in the activity of some other campus unit. Since the record of this event was generated by the archives, it will be added to our collection of archives-generated social media records. We plan to create one record in our collection database to describe all of the social media records gathered by the Archives. Each harvest project will be described at the file level; if the collection grows over time, files could be grouped at a series level by topic or some other criteria. We plan to publish datasets through our digital library (<https://digital.library.illinois.edu>) and also possibly through the DocNow Catalog. GW Libraries has also provided some good thoughts about releasing datasets through their blog.<sup>41</sup>

---

<sup>40</sup> In the case of student organization, archivists should gain permission (preferably written via a deed of gift or other instrument) from a representative of the student organization before beginning harvesting, since it is unlikely that the university or college will own copyright on the harvested data.

<sup>41</sup> Littman, "Releasing Datasets."

<https://gwu-libraries.github.io/sfm-ui/posts/2017-03-15-releasing-datasets>.



### *Collecting Topical or Subject-Based Records*

Archives, of course, tend to have particular subject strengths, often reflecting the interests or characteristics of the parent institution. While considerable discretion is in order, an archives may wish to gather records that support future research in a general area. In fact, GW Libraries uses the tool in just this way, as the described in Justin Littman's post about a single day of collection activity.<sup>42</sup> At that time, GW Libraries was building collections about Congress, Healthcare, "Make America Great Again," the Donald Trump Administration opposition the Administration, and several topics of interest to GWU faculty. Obviously, subject-based collection efforts could become rather amorphous or result in huge datasets, if not properly scoped. Nevertheless, this strategy might be employed in a similar way to event-based collecting, with one difference: It typically would make most sense to employ a twitter filter harvest rather than a search harvest, because the appropriate search terms are known ahead of time. Otherwise, the same considerations mentioned for event collecting apply to topical collecting.

Ideally, collection efforts for subject or topical collections will be coordinated among multiple partners, as more archives being using SFM. For example, GW Libraries is already collecting all tweets by Illinois members of the US House of Representatives and Senate, as they are for all states. State repositories may have an interest in developing complementary (e.g. state-level) collections to complement the larger datasets—for example tweets about or in response to politicians representing the states, not just tweets from the representatives and senators. In this sense, 'person-based' collecting could be seen as a variant of subject-based collecting, where you are more interested in tweets about a person, than by that person.

### **Research Consultation Service**

As should be obvious from the foregoing discussion, social media collection is not a 'set it and forget it' function, no more so than other web archiving efforts. Even in the simplest possible implementation scenarios, library and archives staff will need to develop an appropriate policy framework and will need to monitor and actively manage the service.

The need of active management is even more true if the collecting organization intends to support end user captures, to complement its own collecting activities. This use of the tool seems very likely in many repositories. Partnerships between archivists, students, and faculty will be necessary for both groups to get the most value from SFM. Archivists can help students and scholars understand how the application work, explain privacy issues, help them seed collections, review whether the tool is collecting the data they need, and help them package it for potential export and use. In turn, users can help identify collection areas and shape tool implementation.

---

<sup>42</sup> Littman, "A Day of Collecting," <https://gwu-libraries.github.io/sfm-ui/posts/2017-05-08-day-of-collecting>.

In fact, GW Libraries originally built the service to meet faculty and student needs. With a strong campus focus on public policy/administration, media studies, politics, and public health, many GW faculty and students had interest in monitoring, capturing, and then analyzing social media data for potential insight into particular research questions. GW Libraries consult with potential users, help them registering to use the tool, and let them collecting social media records, which users the export for analysis. Other repositories could do likewise, with three projected outcomes, in order of increasing complexity and support demands:

### *Researcher Collects, Uses, and Discards*

In this scenario, the Library would simply act as facilitator. A researcher creates their own collections for their own use. Data stays in SFM as long needed for the current project and is exported from the UI by the user or by an administrator from the command line. When the project is over, the data is then is deactivated or deleted by SFM staff—a function that is available only to administrators of the system and is completed through the database.

### *Researcher Collects and Library Preserves*

A researcher creates a dataset for their own use, but makes an agreement that the archives will accession and preserve it for its potential future research value. In this case, a specific agreement to donate or transfer should be made with the collecting researcher. This could be via a simple email but ideally would be formalized in deed of gift or other legal instrument, in which the researcher gives whatever right he or she has in the data to the repository.<sup>43</sup> Alternatively, if the donor does not wish to donate the dataset, the archives could be provided a license to distribute it or make copies available, either immediately or after a set period of time. Such an arrangement could be suitable in the case of a collector who does not wish to immediately share data immediately, to protect a forthcoming research work, but who wishes to share it longer-term.

### *Library/Archives Creates and Preserves*

In this case, the archives or library would directly manage the development of new collection sets, collections, and seeds, seeking to develop research resources that need an immediate need, but which also might have some longer term research value. A strategy like this could complement the building of subject-based collections mentioned in the last section. It would build partnerships with Library user to identify and capture collections with near and potentially long-term research use. In this respect, students and faculty members can help archives identify

---

<sup>43</sup> Copyright in collected by SFM is governed by the social media providers' terms of service, so the rights transferred by this deed will be limited, if not negligible, so far as they concern the actual data. For example, in the case of tweets, the copyright of the original tweet lies with its original creator. But the deed would at least specify the the donor's desire to see the results of the collecting activity preserved and made accessible to other scholars. This is an important factor since datasets themselves have separate intellectual property rights from the underlying data. For further discussion and advice, see Cornell University, "Intellectual Property Rights in Data Management" and Carroll, "Sharing Research Data and Intellectual Property Law: A Primer."

relevant collecting areas and shape uses of the tool. For example, users may be able to provide some direction concerning tool development and access methods. As Ian Milligan has noted, future historians will need to develop text mining and manipulation skills to interpret the voluminous records generated by web archiving tools.<sup>44</sup> As additional archives work with SFM, users can help shape future version of the tools, in particular any access methods that might be developed to enhance for SFM data.

When an archives or library is leading the collecting efforts, several potential approaches suggest themselves. For example, a subject librarian may be co-teaching or providing support services for a faculty member, say a professor who specialises in health care economics. When the students are assigned a project on consumer attitudes toward drug prices, emergency room costs, or some other topic, the faculty member and subject librarian consult with each other and decide to create a dataset that (hopefully) speaks to these topics. After collection has been completed, they export some datasets for the student groups to analyze. If the data provides useful, it could be retained for future use, or the harvesting could be extended over one or more years.

As another example, consider an archivist who is supporting a graduate history seminar about protest movements past and present. After reading widely in the secondary literature, the students are asked to examine and reporting on primary source datasets. Referred by their professor, student meets with the archivist and asks if they could collect social media data relating to protest groups on opposite sides of a campus issue. The archivist suggests some potential twitter search and filter harvests, then configures SFM to collect the data. Provided a system login, students download and analyze datasets for their class paper, and after the course is over, the archivist continues to harvest the data for potential future use.

As a final example, consider the case of an institution that holds strong library collections related to a particularly country, such as the Philippines, including books, newspapers, and periodicals in Filipino, English, Spanish, and other languages. In addition, the Library's special collections department has acquired well-known and heavily-used manuscript collections from Filipino immigrant groups. To support faculty and student research, the library began collecting web resources in 2008, and it now wishes to supplement those collecting efforts by capturing social media, relating to both the Philippines proper and the local immigrants. SFM provides Library staff the means to do so, and even a tool to foster the local community's digital self-archiving.

Obviously, there are many other potential uses to which SFM can be put within a particular archives. While the foregoing discussion mentions some general possibilities, the collecting organization will need to consider carefully which uses it wishes to pursue, translate those uses into specific goals, and then develop policies and procedures to ensure that the activities rest on a solid foundation.

---

<sup>44</sup> Milligan, "The Promise of WebARChive Files."

## Preservation of and Access to Social Media Records

Since so much of daily life—what in future years we will call history—plays out via social media, many libraries and archives that use SFM should aim high. They should seek to preserving of social media records so that they can accessed ten, twenty or even one hundred years hence. Doing this will require considerable long-term planning and some local policy development, not least because social media collecting is a new area, and one which is not fully defined or agreed upon in either the law nor in professional practices. That said, the SFM project staff have provided many resources that will allow repositories and the archival profession to move forward and to advance public debate around these topics, while taking practical steps to actually preserve the records SFM captures.

### Establishing a Policy Basis for Access

As previously mentioned, the SFM team has developed and released social media collection development guidelines. Rather than offering prescriptive advice or specific policy recommendations, the guidelines introduce a set of questions that archives and libraries can use in guiding conversations about how to capture, preserve, and provide access to social media data, focusing on the following areas:

- Ethics
- API Terms of Service
- Harvest Scoping
- Documenting collecting decisions
- Access

Reading through the guidance and thinking about the various issues they raise, it may seem easy to become paralyzed. But the questions can and should be framed in light of overall archival objectives and with an understanding that risks can be mitigated by developing, implementing, and documenting policies and procedures that govern the way social media records will be accessed. In other words, by thinking through the end goals, repositories will shape and be able to preserve collections that meet projected access needs.

To help archives begin the process of developing these policies, I would like to describe three basic access scenarios and list some potential use cases for each one. These scenarios are not prescriptive. Repositories may wish to pursue strategies that are less (or more) risk averse, in line with local interpretations of the questions discussed in SFM's collection development guidelines. But here are some examples of potential access scenarios:

*Scenario One:* Archives preserves and provides access ONLY to tweet ids. In this scenario, not much is being preserved. Twitter IDs have no inherent meaning. Data preservation and the ability to contextualize tweets depend wholly on Twitter. This scenario would be useful if the collecting repository cared only about the current use of the data. Such a strategy might also be useful if the collecting organization is harvesting data that implicates the privacy rights of many third parties, such as a repository seeking to document health issues, drug abuse, domestic violence, political resistance in an authoritarian country, or some other topic where there is a potential risk to the safety or privacy of individuals who have not consented to have information about the collected and distributed.

*Scenario Two:* In other cases, archives may wish to preserve and provides access to the full content returned by the API, but allow access to the full data only under very strict controls, such as signed agreement to a condition of use form. Such a scenario may be warranted when a large number of ‘third party’ social media records have been collected and when they speak to a topic of public importance, but the records are perceived as bearing risk to the institution, should deleted social media records be distributed or should the privacy of third parties be compromised. In this case, the signed condition of use form would inform the user of their responsibilities to use the social media records in a way that meets legal, institutional, and ethical requirements.

*Scenario Three:* In many cases, a collecting repository will want capture and preserve the full complement of data the social media service provider supplies, post tweet ids on line, and provide on-request access to full data under the terms of a use policy. Researchers would access social media records in a search room or remotely after being provided a copy by the archivist. The archivist would not closely review the content of the tweets, take steps to remove deleted tweets, or require that the user sign an condition of use form. Such a scenario might be particularly warranted when it is known that all of the tweets in the dataset were generated by the parent institution of the archives that collected them or where copyright to the tweets has been granted to the organization by the owner of the underlying copyrights. In this case, there is low risk to the repository. Even if a tweet has been deleted and is unavailable from the service provider, the underlying copyright is owned by the institution. Such a scenario might also be warranted in cases where the tweets involve prominent public figures whose activities via twitter and other social media services speak to their public role, for example, politicians or celebrities.

## **Export Options and Scenarios**

Whatever access scenario an archives envisions, there are two ways that information can be moved out of SFM: (1) by preparing exports in the user interface or via the command line and (2) by copying collection sets or collections from the server to a new location. Either method could play a role in a preservation strategy, and archival organizations will want to consider which method best suits their needs as well as anticipated uses to which the datasets will be put. Accordingly, this section of the report reviews options before listing factors to consider when developing an export, packaging, and preservation strategy.

## Exporting Data through the User Interface

One of the most useful features of SFM is the ability to export datasets directly from the user interface and (for those who have access) from the terminal. The main purpose of this exporter is to facilitate research access to the harvested resources, so that they can be browsed, parsed, mined, visualized or otherwise manipulated with a tool of the user's choice. That said, it must be noted carefully that the export feature currently **ONLY** exports the social media data itself, not profile images, embedded media, or linked webresources.

As described in the documentation and as shown in figure eighteen, a user simply navigates to a collection, selects appropriate export parameters (seeds, harvest and post dates), then initiates an export. When the export process has finished, SFM emails the user, who then downloads a zip file from application's export tab. The export includes a readme file alongside the data, which is aggregated into one or more JSON files. A copy of the files is also stored on the server, where users can access it through the terminal or file system, if desired.

**Request Export**  
If no seeds are checked, all seeds will be exported.

**Seeds**

- Token: uofiadmissions; Uid: 26793352
- Token: Illinois\_Alma; Uid: 18165866
- Token: IlliniAthletics; Uid: 295948577

**Export format\***  
Full JSON

**Export segment size**  
Single file

Number of items per file.

Deduplicate

**Item date start**  
[ ]

**Item date end**  
[ ]

**Harvest date start**  
[ ]

**Harvest date end**  
[ ]

Collecting and using data from social media platforms is subject to those platforms' terms (Twitter, Flickr, Sina Weibo, Tumblr), as you agreed to them when you created your social media account. Social Feed Manager respects those platforms' terms as an application (Twitter, Flickr, Sina Weibo, Tumblr).

Social Feed Manager provides data to you for your research and academic use. Social media platforms' terms of service **generally do not allow republishing of full datasets**, and you should refer to their terms to understand what you may share. Authors typically retain rights and ownership to their content.

In addition to respecting the platforms' terms, as a user of Social Feed Manager and data collected within it, it is your responsibility to consider the ethical aspects of collecting and using social media data. Your discipline or professional organization may offer guidance. Here are [a few resources to consider](#).

**Selected seeds:**

- Token: uofiadmissions; Uid: 26793352

**Id:** facab4eb3dc044428929db6fe733781e  
**Requested:** May 3, 2017, 7:14:15 p.m. CDT  
**Ended:** May 3, 2017, 7:14:23 p.m. CDT  
**Status:** Success  
**Performed by:** Twitter Rest Exporter on 368daebda7d1 (14)

**Export type:** twitter\_user\_timeline  
**Format:** json\_full  
**Export segment size:** Single file  
**Deduplicate:** True  
**Item start date:** None  
**Item end date:** None  
**Harvest start date:** None  
**Harvest end date:** None

Filename
README.txt
facab4eb3dc044428929db6fe733781e_001.json

**README (2).txt**

This is an export created with Social Feed Manager.

**EXPORT INFORMATION**

Selected seeds:

- Token: uofiadmissions; Uid: 26793352

Export id: facab4eb3dc044428929db6fe733781e  
Export type: twitter\_user\_timeline  
Format: Full JSON  
Export completed: May 3, 2017, 7:14:23 p.m. CDT  
Deduplicated: Yes

**COLLECTION INFORMATION**

Collection name: University-wide Twitter Account Timelines  
Collection id: 063127f8b2914793980aee0ef3b3dc  
Collection set: University of Illinois Accounts (collection set id 0fa3938f39904bc1890b3c373a7c0e0e)  
Harvest type: twitter\_user\_timeline  
Schedule: One time harvest

**Harvest options:**

- Media: No
- Incremental: Yes
- Web resources: No
- User images: No

**Seeds:**

- Token: uofiadmissions; Uid: 26793352 - Active
- Token: Illinois\_Alma; Uid: 18165866 - Active
- Token: IlliniAthletics; Uid: 295948577 - Active

**Change Log:**

Change to University-wide Twitter Account Timelines (collection) on May 3, 2017, 5:12:14 p.m. CDT by orion:

- is\_active: "True" changed to "False"

**Figure Eighteen: Full JSON Export via User Interface**

A few other features of the export feature should be noted carefully, when considering how exports might be ingested and preserved in a digital repository:

- SFM supports several different types of exports. The full JSON response from the social

media provider can be exported, or the user can request several derivative formats, including limited JSON, comma separated values, or even just the social media ID's (dehydrated tweets).

- The documentation provides a good overview of the formats, and the data dictionaries referenced there prove quite useful in understanding the contents of the various export files files, useful to both end users and staff storing SFM data for the long term.<sup>45</sup> Since the data dictionaries are subject to change over time, repositories may need to check the mirrored copies of them in the Internet Archives to understand historical datasets being exported or preserved.
- Users and archives' staff should note that JSON exports are provided in as line-oriented JSON—as supplied by the social media service providers—not as complete JSON objects. In other words they are not wrapped in an opening or closing bracket nor are the root JSON objects separated by commas. The readme files point to one or more JSON files, which can then be read with a line oriented JSON processor, one JSON object per line.
- It is very useful that SFM provides an export of the dehydrated tweets, since that gives repositories a data object that can be published under Twitter's terms of service. Datasets can be provided either on a archival website or through a resource like Documenting the Now's catalog of tweet IDs.<sup>46</sup>

From the end user's and the archivist's perspective, each derivative format may meet a unique research need. For this reason, archives may wish to leave an "access master" of the data in SFM so that particular datasets can be supplied on demand. Using SFM as an access master holds a distinct advantage: since the exporter pulls data from multiple WARC files, then aggregates it into JSON or text files, the social media data would not need to be extracted from a preservation repository or a compressed, serialized WARC files, before it is supplied to a user.

The distribution of captured and stored social media data is limited by the terms of service for a particular API's (as the SFM software very clearly and helpfully notes on the export page, even linking to some resources that can inform user actions). So if an archives chooses to pursue this strategy of keeping data in SFM as an 'access master,' staff will need to carefully administer user accounts in line with a local policy. For example, they should ensure that end users are provided access only to the data that they need and that they are informed of policy requirements and ethical considerations concerning the use of social media data.

Aside from allowing the export of access copies, SFM's export feature has another potential use: Repositories may wish to preserve exported datasets in a digital repository. For example, SFM might be used to generate dissemination packets of data, which could then be stored in an external preservation repository system. The dehydrated tweet IDs could be provided an 'online' dissemination packet (through the web), the various JSON and csv exports in a

---

<sup>45</sup> [http://sfm.readthedocs.io/en/latest/data\\_dictionary.html](http://sfm.readthedocs.io/en/latest/data_dictionary.html).

<sup>46</sup> <http://www.docnow.io/catalog/>.

'nearline' packet (available only after mediation by an archivists), and the full collection folder from the server (with all of the mirrored metadata from the database and full JSON WARC and the Webfiles WARC could serve as the Archival Information Packet. This topic is discussed in more detail in the packaging section below,.

### *Copying Datasets from the Server*

It is easy to copy data from SFM's data store to another location, provided that a user has direct access to the filesystem or docker volume where the data is stored. As explained in the documentation, administrators can execute a recursive copy command against a folder storing social media records and associated content (such as linked web pages and embedded images). Collections, collection sets, or even all of the data in an entire SFM instance can be moved about in this way. For example, `cp -r sfm-data/collection_set/b385d990cc224a39a66ae8e5392ac0dc /target_folder_location/` would copy an entire collection set to the specified target folder. Other linux commands, such as scp (secure file copy) could be used to move data to other computers on a network. If the storage volume is made available in other ways (say, via the sftp or smb protocols) it could be accessed using other tools, such as graphical file managers like Windows Explorer or the Mac Finder.<sup>47</sup> In addition, SFM has the ability to import collection sets and collections. So, if a collection has been copied outside of SFM it can always be placed back into its original environment. This might be useful if, for instance, a future archivist wishes to prepare a new data export to meet some unanticipated user need.

Regardless of the specific methods by which a user accesses SFM's data, **everything** in an SFM instance—including social media records and metadata that is stored in the database—is included in the `collection_set` folders (as previously noted and as illustrated in figures five and six). Thus, the collection sets compose a consistently structured and complete set of digital objects. In other words, SFM's collection data includes many the elements that might make up a submission information packet, something that can be prepared for ingest into a preservation repository. The basic point to be borne in mind here is that the collection data on the server includes much more metadata about the collections than the exports do, so that collection data is much more suitable as an archival information packets than the exports, which are more akin to dissemination information packets, in OAIS terms.

### **Packaging Recommendations and Descriptive Metadata**

One important thing most archives will want to do is ingest social media data into a preservation repository. Since SFM is not a preservation service and since many archives and libraries are developing separate preservation repositories, it will be ideal if both the social media data and associated information (such as captured media and linked webpages) can be moved to other

---

<sup>47</sup> If a repository uses docker volumes for storage other commands would be used, since the files would not be available using typical file system operations. The specific methods to access docker volumes are explained more fully in the documentation.



systems. In looking at SFM's technical structure (relationship of database records to WARC's), technical documentation, and exports, most of the pieces are in place to allow for a fulsome capture and preservation of information outside the system, including the ability to package data, then describe it in an external system. Regardless of the specific systems and tools used or the specific description decisions that are made, packaging work could take place in three distinct ways: via export based packaging, via collection-based packaging, or via a combination method.

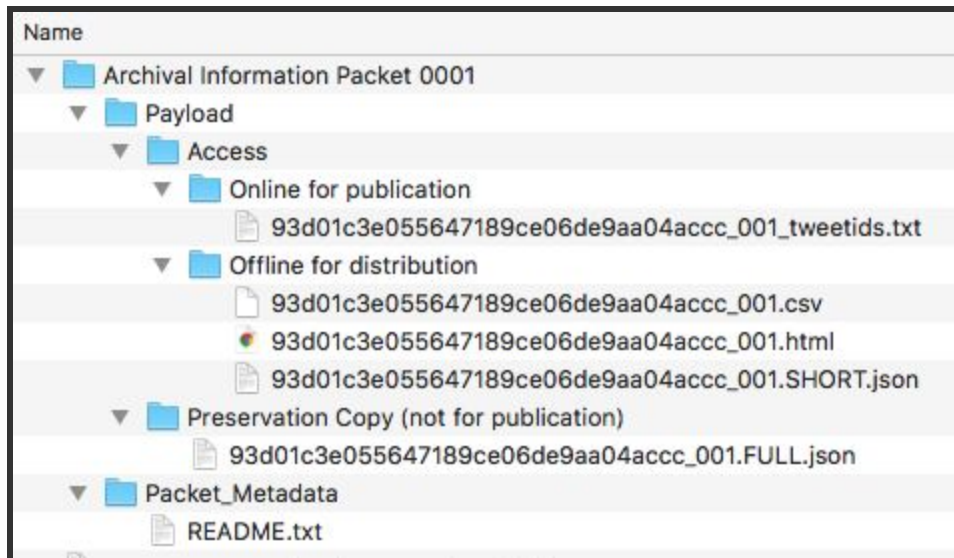
Accordingly, this section of the report describes those three options, before providing some description recommendations that apply regardless of the arrangement and packaging scheme that has been used. It concludes with a short discussion of accruals and pruning SFM

### *Export-Based Packaging*

In the export-based packaging scenario, the archives would commit to preserving only social media data, **NOT** web resources, embedded media, and metadata that SFM has generated about the captures. This approach would not be suitable for archives that wish to preserve web resources or embedded files, and it would sacrifice a large amount of metadata and provenance information the SFM generates in the course of capturing data. Those provisos aside, it would provide the archives a relatively easy-to-preserve object in ASCII text formats, both for the metadata and the harvested resources.

Most likely, repositories would choose to use export-based packaging if they are more interested in preserving the data rather than a record of their actions in capturing it. For example, if an archives collects tweets around a specific event, the archivist or collector could take the following steps to finalize a preservation object once data collection has been completed:

- Turn harvesting off at the end of the project
- Export readme and files for various dissemination packets:
  - For online access: tweetids.
  - For 'nearline' access: csv files, html files or json files. As defined by the archives' policy and procedures, access to these could be provided in a controlled fashion, for example in the archives/library reference room or a limited range of IP addresses.
- Export readme and full json files for 'preservation'/archival information packet.
- Copy the readme and json files a locally defined package structure (see figure nineteen for an example) and name the files appropriately to reflect their content
- Ingest them into a preservation repository
- Describe them using locally-defined descriptive tools.



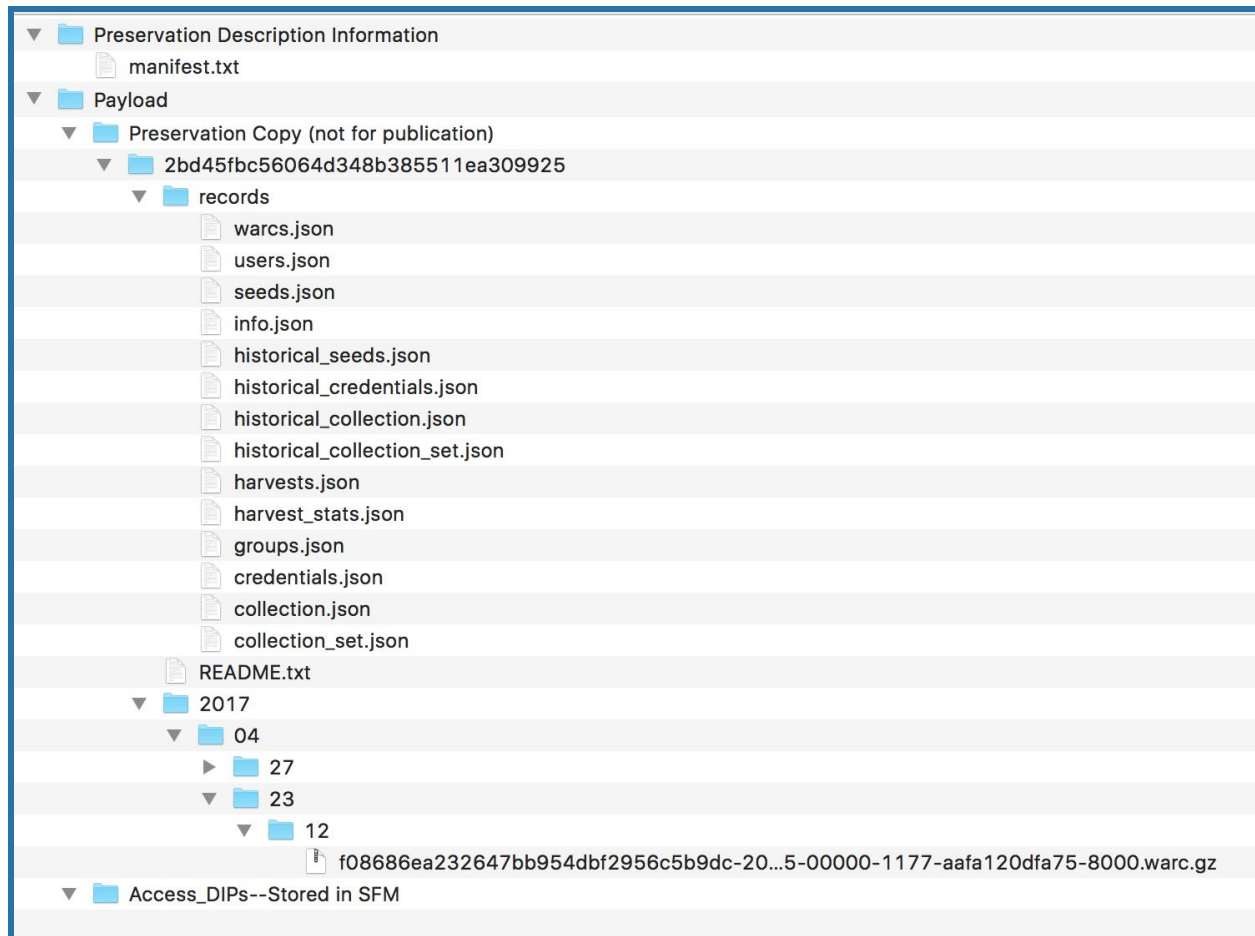
**Figure Nineteen: Example Package Structure for SFM data (Non-normative) Using Exported Data**

### *Collection Set or Collection-Based Packaging*

Instead of packaging and preserving exports of SFM data, repositories may wish to package and preserve collection sets or collections, then ingest them into a preservation repository. This strategy would be particularly appropriate in cases where a copy of the collection set or collection data remains in SFM as an ‘access master,’ from which dissemination copies could be exported or produced on an as-needed basis, in response to specific user requests.

In this case, the packaging routine would be relatively simple:

- Turn harvesting off at the end of the project
- Copy the full contents of the appropriate collections set or collection folder from from the sfm-data folder to an external storage location (include the Readme file, the various JSON files, and the WARC in their original folder structure)
- Place all of these files into a locally-defined package structure (see figure twenty for an example)
- Ingest them into a preservation repository
- Use locally-defined tools to describe (a) the files in the preservation repository; and (b) the export options available from the access master copy in the SFM instance.

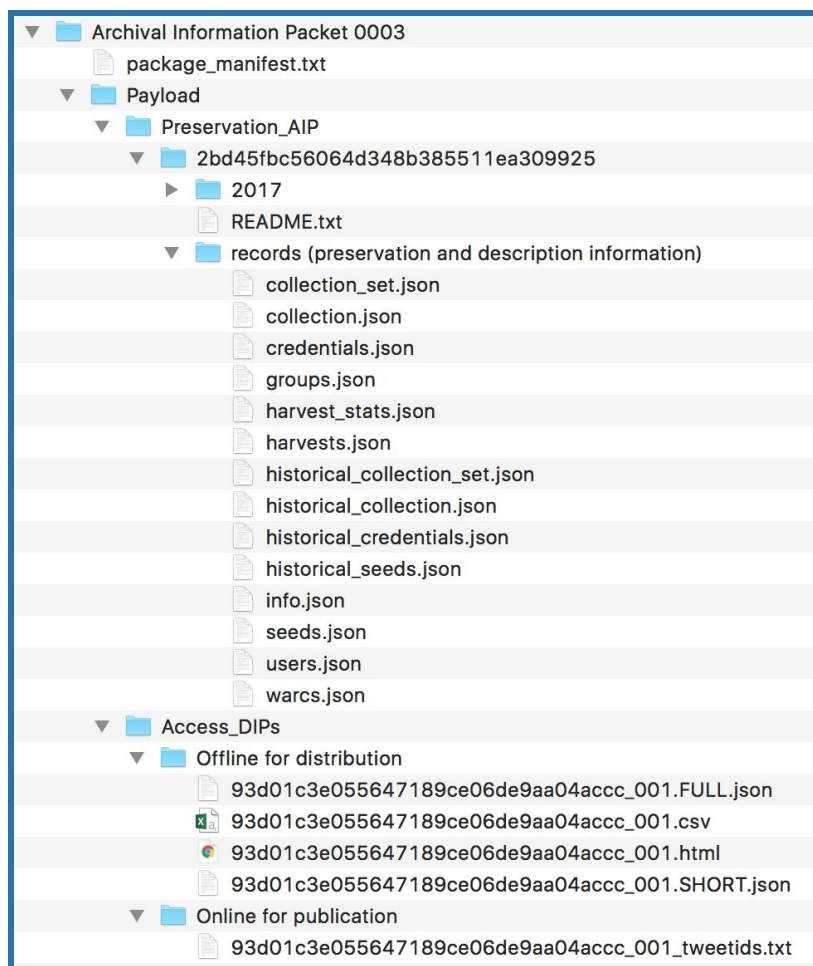


**Figure Twenty: Example Collection-based Packet Structure (Non-normative)**

This approach holds some advantages, most notably the fact that access copies could be generated at any time of out of SFM. In addition, all of the metadata stored in SFM has also been included in the packet in JSON formation, making for a relatively self-describing packet. However, collecting organizations should monitor a few points. First, data integrity will need to be monitored very closely. Not only is binary data (not simply text files) being stored, but those files are stored as a bitstream (in WARCs) and compressed to the gzip format. As a result, it will be wise to undertake additional preservation planning. For example, if repositories are committing to preserve and provide access to embedded media or web content, format monitoring and possibly migration will need to be considered. In addition, repositories choosing this approach should be aware that there are committing to a rather complex package structure. In the case of long-term or frequently harvests, many warc files will be generated, and they will be nested in a complex data-based folder structure, which could potentially complicate future access, should it become necessary to access data from a preservation repository, instead of directly from the 'access-master' in SFM.

### Combination Packaging

A final approach to consider is what might be called combination packaging, that is to say, including both exported data and collection data in a single package to be preserved outside of SFM. In essence, the exported data would be treated as an access master and the collection data as a preservation master, as shown in figure twenty-one. This would work best in cases where all collection data is exported, i.e. where no seed selections or data limits are imposed; otherwise the exports and the collection data would be out of sync with each other. It would also be most suitable in cases where the collecting organization does not plan to keep data in SFM and wishes to store the data in some kind of unified preservation repository. Data would be accessible with its full provenance in the preservation folder, and copies could be made available from the access folder, without needing to undergo additional transformation by the archives or the user. However, the same requirements for data monitoring and migration would apply as noted in the previous subsection on collection-based packaging.



**Figure Twenty-one: Combination Package with Export and Collection Data from Server**

For example, SFM provides archivists and other users with the ability to record some basic

metadata regarding collection sets and collections, as well as the ability to add change notes whenever an action is undertaken in the system. As a best practice, archivists should use these fields systematically, since they provide a way to record information about a collection set or collection at the time it is known, or to explain their rationale for adding seeds, tweaking harvest parameters, or undertaking other actions. In essence, these notes, along with the set of metadata that SFM generates about its own actions, leave behind a deep and meaningful set of provenance metadata about the records that SFM captures.

### *Recording Descriptive Metadata*

SFM is not a descriptive tool, and it is expected that most archives will describe the SFM collections that they have packaged in an external descriptive system. While many potential metadata schema could be employed, the first and most basic tactic should be to describe an aggregation of social media data, rather than individual social media records. Whether the results of an SFM collecting activities are seen as a ‘collection,’ a ‘series,’ a ‘subseries,’ or a ‘file’ in archival terms, the main need is for an accurate and complete description of the aggregate, so that potential users can discover a set of social media files that may help answer a particular research question.

Well-constructed descriptive records will build trust in the value and authenticity of collecting efforts undertaken with SFM. Accordingly, the following metadata elements are suggested as a minimal set of data points that should be recorded in an external descriptive record about social media data. These are suggested simply to provide archivists the means to record minimal descriptive metadata along the lines suggested by *Describing Archives: A Content Standard* for “Single-Level Required” or “Multi-level Required” Records, but would also have some applicability toward other metadata schema.

### **Figure Twenty-two: Metadata for Describing Collection Sets, Collections, or Exports**

<b>Element</b>	<b>Notes</b>
Inclusive Dates	Human recorded date span for the dates between which social media records in the collection were posted.
Capture Dates	Human recorded date span covering dates that social media were collected or captured, can be taken from SFM database or user interface.
Extent	Machine generated; total size of the archive in bytes, including all various representations of the data as stored in the preservation repository and/or access system.
Scope and Content	Based on description from SFM, describes the nature of the records included, including account owners, people, organizations, or subjects included. May be helpful to includes

	lists of seeds, account names, or hashtags used in search or filter collections or what they mean.
Collecting Party	Name of the Archives or person responsible for the collecting activity
Access and Use Statement	Narrative statement of conditions under which the repository provide access to the materials, and under which they might be used in publications or other ways. Agreement to conditions of use forms or other requirements should be described here.
Custodial History	Note any prior people or organizations that held the social media collection before it was deposited in the archives.
Arrangement Note	Describe the packaging schema used, or the internal structure of the collection (including any search parameters or other user input used to shape the collecting activity, and bearing on the nature of the preserved records.) If exported data is being described, list and relevant export decisions that shaped the nature and arrangement of the materials.
Immediate source of acquisition	List the donor or depositor of materials, or describes conditions under which the archives undertook the collecting activity.
Note	Used to record provenance information or anything else relevant that doesn't fit in the other fields

### *Accruals and Pruning*

For long-term projects, archivists may need to define a strategy for incorporating accruals into externally packaged record sets. For example, the archives might be collecting a timestream for a University social media account and wish to deposit tweets every year. Or perhaps the archives has defined a twitter filter search that generates several gigabytes of data annually. In either case, the archives would add an accrual to the preservation repository.

Generally speaking, easiest solution would just be to replace the entire folder in the preservation repository with the data from SFM (preferably comparing comparing the checksum values of any files that are being replaced), then updating descriptive metadata in the archives' catalog. This strategy makes sense if no data is ever deleted from SFM. And in many cases this will be completely feasible or even preferable, since the cost of pruning or keeping SFM 'clean' may be much higher than simply adding more storage.

But if the archives needs to prune the SFM instance due to disk space limits, policy, or other reasons, it will need to define a different method of adding accruals. Since SFM allows users to

produce time based exports (either by date of posting or date of capture), the user could select a year's worth of social media records, export their metadata, copy the appropriate warc files from the preservation repository, and fit everything into the preservation repository, packaged according to local requirements for adding accruals. This would allow the archives to delete data from the server, while retaining some ability to import it back to SFM if that ever becomes necessary. This would work well for instances where collections are static and are not expected to grow quickly, and this is a workflow that will be familiar to archivists who manage other born-digital content. That said, the processes outlined above require considerable knowledge about SFM by staff, access to the server or storage location where WARCs are kept, and quite a bit of effort. It might also be error prone, and it would be very useful if portions of this workflow could be automated. For example, it would be very helpful if future versions of SFM facilitate the export of an entire preservation object—the JSON and associated warc files from the collection folder—for a user-defined range of capture dates. That would facilitate a more automated means of adding accruals to an external repository.

The following section of this consultation report provides some additional thoughts as to how the current export functionality, as well as other features, of SFM might be enhanced.

## Community Recommendations

Open source projects are most likely to succeed and remain viable when they meet a clear need, when their underlying technologies are well supported, when their own technical model is well designed and documented, and when a user community coalesces around the application. Given these factors, and with NHRPC support coming to a close, developers, archives/libraries, and the funding community are well-placed to enhance SFM with additional functionality and to ensure the viability and sustainability of the tool.

### For Developers

Based on my experience using the tool, GW Libraries constructed it on a very solid and extensible technical foundation. That said, SFM might be enhanced over the upcoming years to become even more useful to the archives and library community. Following completion of the work supported by NHRPC, projected technical enhancements could be considered in the following areas.

#### *Enhance Capture Capabilities*

- *Add tweet rehydrator.* It would be very useful for SFM to support the capture of full tweet metadata from a set of tweet ids. Maybe a user has downloaded a set of tweet IDs from the Documenting the Now Catalog or Dataverse, or has previously captured tweets using another tool, such as the Twitter Archiving Google Sheet.<sup>48</sup> They have the tweet

---

<sup>48</sup>Some tweet id datasets may be downloaded from <http://www.docnow.io/catalog/> and <https://dataverse.harvard.edu/dataverse/gwu-libraries>.

IDs and possibly some metadata, but not the complete JSON from the API. In this case, SFM could be used to reconstruct the dataset. This would require a new harvest type option, in which a user would upload a set of tweet IDs. It would require the integration of the hydrator software developed through the doc now project or the development of new code that can make API calls to the function that gets statuses by ID.

- *Facebook integration:* SFM users and developers might study what it would take to develop a Facebook harvester or harvester, then develop specifications and an extension to SFM. For example, perhaps SFM could be used to capture Facebook data for a user and people he or she follows, like SFM currently collects data from Sina Weibo. But given the difficulties of working with the Facebook API, much work is necessary.
- *Other APIs:* Other social media APIs could also be studied for potential integration.

### *Refine Export and Packaging Options*

- *Export Metadata:* When a user exports data, it would be helpful if metadata about the exported social media posts were serialized from the SFM database into JSON format, then written into the export packet. This would make the package structure for exported data similar to the collection sets and collections, including both the post and metadata about them. It would also make the exported data into a more complete preservation object, including not only the records to be preserved and a read me file, but also preservation metadata.
- *Full Content Exports:* It would also be very helpful if SFM included the full content in exports for collections that include webresources. Then the exports would hold not just social media, but also harvested resources, embedded media, and profile images. This too would make the exports into a more rounded and complete preservation object.
- *Enhanced Time-based Collection Exports:* Similarly, SFM might support the export of more complete packets for time limited exports. For example, if a user were to export all records in a collection for a particular harvest period, copies of the relevant folders from the collection directory on the server might be copied into the export packet. This would facilitate the addition of accruals to an external repository.
- *API Integrations:* Longer term (i.e. in another grant cycle), the team might explore adding an API to facilitate data transfers.<sup>49</sup> For example, SFM could include a method to retrieve the JSON-ified versions of the read me files—which would also point to the line-oriented JSON and the related WARC files. This would then provide a simple REST-based API to retrieve and expose metadata and social media records, leveraging the existing message system to expose data via defined endpoints for collection sets, collections, seeds, warcs, etc. That way, other services could read it and it would be possible to integrate metadata into other services (such as ArchivesSpace) and WARCs into preservation repositories, facilitating the ingest and packaging process.

---

<sup>49</sup> This could be patterned on or at least informed by the WASAPI project, since that is intended to facilitate the exchange of warcs: <https://github.com/WASAPI-Community/data-transfer-apis>.



## For Libraries and Archives

The most immediate opportunity is for more archives and libraries to implement SFM as a production service. Social media collecting is a new type of archival activity, but one that many archives can and should pursue using SFM.

### *Establish Social Media Collecting as Core Archival Activity*

Notably, SFM allows archivists, perhaps for the first time, the ability to intercept, capture, and preserve records at or nearly at the moment of their creation, as a normal part of archival work.<sup>50</sup> As a result, the records it captures contain much more embedded metadata and provenance information than those typically received by archives. This is important, because in the case of most digital records being captured in archives, practices have changed remarkably little from the days of managing paper. Typically, archivists extract data from storage locations at the end of a period of active use, rather than intercepting them at or near the time they are transmitted from one computer to another. Forensic approaches to archives are so popular at least in part because the profession is still wedded (not always by choice) to the life cycle model of capturing records that have been 'set aside' either willfully or incidentally on storage systems. Because collecting records from an API breaks us free from the model, it has the possibility to transform archival work. As more and more services make APIs available, archivists will have greater and greater opportunity to capture records at or near the point of their creation.

Yet of course there is a more significant reason for archivists to collect social media records: They document and shed light upon a changing society. While this may seem self-apparent to the historian or archivist, it needs to be demonstrated to local communities and users of archives.

By capturing, preserving, and providing access to records that are important to users and that connect to their daily lives, we demonstrate not only that we are trusted stewards of old cultural treasures (an interesting but limited role) but active agents who document history as it happens. For this reason, and for so many others, it is critical that archives establish social media collecting as a core archival activity.

### *Advocate for Enhanced Preservation and Access Rights*

The corporations that people use to produce social media records determine how and whether those records can be preserved. This differentiates social media from most other documentary formats, and the archives/library community has a stake in demonstrating the value of what we can do to preserve the records that people produce while using their technologies. Ideally, we

---

<sup>50</sup> In this respect, SFM puts into practice recommendations such as those made long ago by David Bearman, that archivists capture or 'trap' records at the time they are transmitted, encapsulating all of the system metadata that allows them to be subjected to further manipulation. See Section III 1, "Capture" of Bearman, "Item Level Control."

would benefit from changes to copyright law, granting libraries and archives special exceptions to capture, preserve, and provide access to social media records. Yet we should not count on the success of efforts to change the law, which are costly, time consuming, and uncertain to succeed.<sup>51</sup>

In the meantime, there is much work to be done. The more frequently that archives implement SFM, the more persuasively that our community will demonstrate the value of capturing social media. By making our interests and value known, we are more likely to attract the attention of social media service providers, whose current terms of service make it difficult for libraries and archives to capture and preserve social media records.

As this report was being finalized, the archives/library community demonstrated its ability to influence policy. On May 17th, Twitter released an update to its terms of service, an update that seemingly restricted the ability of archives to distribute tweet id datasets. But after listening to the community, Twitter issued an update/clarification, granting 'researchers' the ability to distribute tweet ID datasets of unlimited size. Justin Littman provided an excellent description of this incident on the SFM project blog, concluding: "It is evident that Twitter was listening and does value the academic research that is done with Twitter data. It further highlights the need for our communities to proactively engage with Twitter as a "good partner."<sup>52</sup> When more archives work with social media data, perhaps other social media providers, such as Facebook, will also begin listening to libraries and archives. While we have a long way to go, the community can find a great deal of value when we bring our needs to the attention of social media companies, since they ultimately control the API's and tools that make social media archiving possible. They can only respond to our requests if we present them fairly and responsibly. It would be especially valuable if the archives and library community were provided an opportunity to discuss social media preservation with Facebook.

## **For the Community and Funding Partners**

As more archives implement social media collecting, they will open possibilities for collaborative collection building and tool development. As noted earlier, a group of repositories could

---

<sup>51</sup> Since 2010, William Maher has been representing the Society of American Archivists at meetings of the World Intellectual Property Organization's Standing Committee on Copyright and Related Rights, seeking to secure library and archives exceptions in the international conventions and treaties that influence national copyright laws. The specific changes he is seeking would enhance the ability of libraries and archives to provide enhanced access to unpublished by in-copyright archival materials. His most recent statement is available at <http://www2.archivists.org/sites/all/files/Society-American-Archivists-Statement-SCCR34.pdf>. His experience demonstrates that changes can only be secured through persistent, tireless advocacy, and any attempt to loosen copyright law regarding social media would be even more bruising and even less certain to succeed, given the relative currency of most social media data and the interests of other parties, such as those advocating for the right to be forgotten.

<sup>52</sup> Littman, "Implications of Changes,"

<https://gwu-libraries.github.io/sfm-ui/posts/2017-05-18-twitter-policy-change>

implement SFM and decide on local collecting priorities to complement an overall documentary object that the group as a whole is pursuing.

One potential barrier to this collaboration is the fact that SFM, at this stage in its evolution, must be installed and managed locally. Technically, the tool is poised for additional technical development and testing. It could evolve into a general social media archiving service that is available to many institutions, similar to the type of service that Archive It currently supplies for websites. Scale testing will help set the stage for further development, but it is far from the only step that will need to be taken if SFM is to become an expected part of archival work, a status that standard web archiving has achieved.

The experience of projects such as SFM and DocNow indicates that archives and libraries are stronger when they collaborate with each other and also with the the government agencies and private foundations that support innovative tools and services. As indicated above, SFM is a prime example of library innovation, and additional technical and policy development would deepen its effectiveness in helping document what people say and do using social media. This is a key task if we are to facilitate deeper understanding of social, political, artistic, medical, religious, and other trends so richly represented in social media services.

Like many relatively young open-source projects, SFM is closely associated with the institution that conceived it: George Washington University Libraries. Most development to date has been completed either by GW Libraries staff or by people employed by the Library. Hopefully, GW Libraries can continue to guide SFM's future development and growth. But the software will most likely remain vibrant when additional partners take an active role in its implementation and development. Libraries and archives must obviously do their part in adopting the tool and, if they are able, contributing code to its development. Policy work is also essential. For example, libraries and archives could collaborate to develop standards for social media preservation and data sharing, or they could collectively bring their needs to social media service providers.

Continued partnerships with funders, such as federal agencies, private foundations, and perhaps even social media service providers will be essential if the community is to advance these efforts. The experience of other open source development projects should be a guide: Projects crucial to the archival mission, such as ArchivesSpace and EPADD, started small, yet grew in stature and influence with the help of granting agencies.

SFM, which is still early in its evolution, could benefit from continued support from funding agencies, particularly in the following three areas:

- Community development and sustainability planning
- Additional technical development, such as that described above
- Policy work, such as establishing a working group of library/archives professionals and representatives from social media companies to develop recommendations or services to facilitate API-based web archiving of social media content.

## Conclusion

Many aspects of modern life are communicated via social media technologies. If archives, libraries, and other cultural heritage organizations are going to acquire and preserve a record of that life, then make it amenable to analysis, they will need to do it using tools like SFM, which captures a more-or-less complete record of social media traffic using the APIs provided by social media companies. Social media archiving takes place at a complex intersection of technical, legal, policy, ethical, the economic concerns. It poses many challenges and risks, but also many potential opportunities and rewards. By implementing SFM and by building experience with social media collecting, archives and libraries will not only build collections that represent modern life, they will shape a historical record that can be mined for generations to come.

# Appendices

## Appendix 1. Reference List/Further Readings

Bearman, David. "Item Level Control and Electronic Recordkeeping." *Archives & Museum Informatics* 10 (1996): 195–245. <http://www.archimuse.com/papers/nhprc/item-lvl.html>, captured at <https://perma.cc/5SSZ-SDTT>.

Carroll, Michael W. "Sharing Research Data and Intellectual Property Law: A Primer." *PLOS Biology* 13, no. 8 (August 27, 2015): e1002235. doi:10.1371/journal.pbio.1002235.

Cornell University Research Data Management Service Group. "Introduction to Intellectual Property Rights in Data Mangement," <https://data.research.cornell.edu/content/intellectual-property>, captured at <https://perma.cc/35MW-BHRW>.

George Washington University Library. "Building Social Media Archives: Collection Development Guidelines." *Social Feed Manager*. Accessed June 5, 2017. <https://gwu-libraries.github.io/sfm-ui/resources/guidelines>.

George Washington University Libraries. *Social Feed Manager (SFM) Documentation*. Accessed June 4, 2017. <https://sfm.readthedocs.io/en/latest/>.

George Washington University Libraries. *Social Feed Manager Project Site*. Accessed June 4, 2017. <https://gwu-libraries.github.io/sfm-ui/>.

Greenwood, Shannon, Andrew Perrin, and Maeve Duggan. "Social Media Update 2016." Pew Research Center: Internet, Science & Tech, November 11, 2016. <http://www.pewinternet.org/2016/11/11/social-media-update-2016/>, captured at <https://perma.cc/53C7-53V9>.

Hawksey, Martin. "TAGS: Twitter Archiving Google Sheet Website." Accessed June 4, 2017. <https://tags.hawksey.info/>, captured at <https://perma.cc/PFB4-6B4Z>.

Kelly, Mat. "Web Archiving Integration Layer (WAIL)." Accessed June 4, 2017. <https://machawk1.github.io/wail/>.

Kreymer, Ilya. *Webrecorderplayer-Electron: Webrecorder Player Code Repository*. 2017. <https://github.com/webrecorder/webrecorderplayer-electron>.

- Kerchner, Daniel, Justin Littman, Christie Peterson, Vakil Smallen, Rachel Trent, and Laura Wrubel. "The Provenance of a Tweet." 2016.  
<https://scholarspace.library.gwu.edu/files/h128nd689>.
- Littman, Justin, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. "API-Based Social Media Collecting as a Form of Web Archiving." *International Journal on Digital Libraries*, December 28, 2016, 1–18.  
doi:10.1007/s00799-016-0201-7.
- Littman, Justin. "A Day of Collecting with Social Feed Manager." *Social Feed Manager*, May 8, 2017. <https://gwu-libraries.github.io/sfm-ui/posts/2017-05-08-day-of-collecting>, captured at <https://perma.cc/G4AJ-MCXM>.
- Littman, Justin. "Implications of Changes in Twitter's Developer Policy." *Social Feed Manager Website*, Updated May 18, 2017.  
<https://gwu-libraries.github.io/sfm-ui/posts/2017-05-18-twitter-policy-change>, captured at <https://perma.cc/HNQ3-5ZY7>.
- Littman, Justin. "Releasing Datasets to Dataverse." *Social Feed Manager*. Accessed June 5, 2017. <https://gwu-libraries.github.io/sfm-ui/posts/2017-03-15-releasing-datasets>, captured at <https://perma.cc/3VCL-M924>.
- Douglas, Jennifer. "Toward More Honest Description." *The American Archivist* 79, no. 1 (June 2016): 26-55. doi:10.17723/0360-9081.79.1.26.
- McFarlane, Greg. "How Facebook, Twitter, Social Media Make Money From You." Investopedia, March 21, 2014.  
<http://www.investopedia.com/stock-analysis/032114/how-facebook-twitter-social-media-make-money-you-twtr-lnkd-fb-goog.aspx>, captured at <https://perma.cc/7MKQ-VAAR>.
- McKay, Aprille C. "Managing Rights and Permissions," in Menzi L. Behrnd-Klodt and Christopher J. Prom, eds., *Rights in the Digital Era*. Trends in Archives Practice Series. Chicago: Society of American Archivists Press, 2015.
- MacNeil, Heather. "Picking Our Text: Archival Description, Authenticity, and the Archivist as Editor." *The American Archivist* 68, no. 2 (September 2005): 247-278.  
doi:10.17723/aarc.68.2.01u65t643570033
- Meehan, Jennifer. "Making the Leap from Parts to Whole: Evidence and Inference in Archival Arrangement and Description." *The American Archivist* 72, no. 1 (April 2009): 72–90.  
doi:10.17723/aarc.72.1.kj672v4907m11x66.
- Miglani, Jitender. "How Facebook Makes Money?" *Revenues & Profits*, June 9, 2015.  
<https://revenuesandprofits.com/how-facebook-makes-money/>, captured at <https://perma.cc/9ZLC-ZLRA>.

- Milligan, Ian. "The Promise of WebARChive Files." *Web Archives for Historians*. January 2015. <https://webarchivehistorians.org/2015/01/22/milligan-presentation-the-promise-of-webarchiv-e-files-exploring-the-internet-archive-as-a-historical-resource/>, captured at <https://perma.cc/Q8Y7-9QJA>.
- Nelson, Theodor H. *Computer Lib; Dream Machines*. 1974. Revised edition Redmond, Wash: Tempus Books of Microsoft Press, 1987.
- Prom, Christopher J. "Installing Social Feed Manager Locally." *Archival Connections*, June 6, 2016. <http://www.archivalconnections.org/installing-social-feed-manager-locally/>, captured at <https://perma.cc/WM9G-9SHY>.
- Tan, Yecheng. "Weibo API Guide." *Social Feed Manager Website*. Updated April 26, 2016. <https://gwu-libraries.github.io/sfm-ui/posts/2016-04-26-weibo-api-guide>, captured at <https://perma.cc/3UHA-44V9>.
- Thomson, Sara Day. "Preserving Social Media." Digital Preservation Coalition, February 15, 2016. <http://dx.doi.org/10.7207/twr16-01>.
- Twitter. "Downloading Your Twitter Archive." *Twitter Help Center*. Accessed June 4, 2017. <https://help.twitter.com/articles/20170160?lang=en>, captured at <https://perma.cc/93BD-BFQU>.
- WASAPI Community. *WASAPI Data Transfer APIs*. Python. 2017. <https://github.com/WASAPI-Community/data-transfer-apis>.

## Appendix 2. Tools and Services for Archiving Social Media

### *ArchiveIT*

- URL: <https://archive-it.org/>
- Overview: Archive-It is a subscription web archiving service from Internet Archive that allows for organizations to harvest, build, and preserve digital content collections. The program is meant to be a user-friendly application that can be used at any time by institutions and their patrons. It offers a web application that allows for archivists to collect, catalog, and manage their collections, as well as full text search capability.
- Use Cases: This tool has the ability to capture any website as long as it is guided to do so. It can be used to capture all social media websites using the Archive-It crawl including Facebook, Flickr, Instagram, Tumblr, Twitter, and Youtube.
- Pros and Cons: While this service does have a price point that must be considered, especially when some options are free, this service does have the full backing of the Internet Archive and their many strong servers that add stability and safety to their partners archives.
- Export Formats: The tool allows for export files by their type, including PDF, image, video, etc.

### *ArchiveSocial*

- URL: <http://archivesocial.com/>
- Overview: ArchiveSocial allows for archiving any social media website with authentic capture that allows for preservation in native format, as well as social media replay, which allows for users to experience the social media as it happens. It allows for 24/7 capture and preservation, advanced search capability, and on demand export.
- Use Cases: This tool can be used to capture social media, its metadata, and the context of the data.
- Pros and Cons: This tool is capable of preserving social media in a way that preserves the data as well as the context of the data and has the ability to archive many social media platforms including Facebook, Twitter, YouTube, LinkedIn, Instagram, Flickr, Pinterest, and Google Plus. While this tool has a cost, there are many cost options that may be able to fit in different types of budgets. It also allows for users to view a free sample of their archive, to see if it would be a fit.
- Export Formats: PDF, Excel, or HTML

### *Hydrator*

- URL: <http://www.docnow.io/>
- Overview: The DocNow Hydrator is a desktop application for hydrating Twitter ID datasets. By providing a set of tweet ids to the application, the JSON response is provided by the API, and written to disk for local analysis.



- Use Cases: Twitter's Terms of Service only allows for datasets of tweet IDs to be shared. The Hydrator helps to turn these tweet IDs back into JSON right from your desktop.
- Pros and Cons: The hydrator provides users a method to harvest twitter datasets to analyze them in a local system, from a set of tweet ID's that have been posted online via twitter's terms of service. However, users should be aware that the response data provided is not necessarily the same as that which was originally posted, and that tweets that have been deleted will not be included. In addition, profile information or other attributes of the message may be different, and the hydrator will not harvest embedded images/videos or webpages.

### *Gnip*

- URL: <https://gnip.com/>
- Overview: Gnip is Twitter's API platform that delivers real-time and historical social media data. Not only does it allow for the collection of data from Twitter, it allows for firehose and managed access to API's of other popular social media sources including Facebook, Instagram, YouTube, etc. This allows for simultaneous collection of social data. It also allows filtered access to the full archive of public tweets.
- Use Cases: Institutions can use this service to purchase tweets that are not available from twitter's public API.
- Pros and Cons: This program allows for the collection of social data across multiple social media accounts. There is, however, there is a limit of 6 accounts that one can link to the program.
- Export Formats: Unclear

### *Lentil*

- URL: <https://github.com/NCSU-Libraries/lentil>
- Overview: This is an open-source program that harvests Instagram images using hashtags. The tagsets are created by administrators and the program pulls the photos, descriptions, and tags into the administrator's dashboard, using the Instagram API.
- Use Cases: Archivists could use this program to develop archives from Instagram, specifically related to the hashtags they wish to target.
- Pros and Cons: The program allowed for hashtags to be targeted, allowing administrators to target specific events and movements happening at an institution and around the world. The program did not, however, have the ability to know if the user it is pulling from is affiliated with the harvesting institution. Even then, it did allow for the administrators to choose to place an image into a collection or not. The software is no longer supported or being developed, and as of June 2016, the API's on which it depended were no longer available to new users, unless registering a specific set of permissions with Instagram, as described at <https://github.com/NCSU-Libraries/Lentil-Instagram-API-submission/>. The long term

prospect of API-based harvesting is uncertain, as Instagram is putting increased focus on its advertising API.

### *Smarsh Archiving Platform*

- URL: <http://www.smarsh.com/archiving-and-compliance>
- Overview: The Smarsh Archiving Platform is one example of a general purpose tool for capturing and managing electronic communications in an organization, including email, instant messaging, text messages, websites, video, and social media. It is collected in its original context, not compressed into a generic text or email format. The system allows for automated capture and export when needed as well as searching by date, keyword, person, etc.
- Use Cases: This platform would be used for institutions to be able to automatically capture all forms of electronic communications, including but not limited to social media, typically in order to meet record retention or legal compliance needs..
- Pros and Cons: Smarsh is ready to use and allows for the collection of data across multiple platforms and has many benefits including automatic capture and searching capabilities. However, the cost of the platform may be seen as a con for many institutions.
- Export Formats: Export is available in many formats.

### *Twarc*

- URL: <https://github.com/DocNow/twarc>
- Overview: Twarc is a command line tool and Python library that archives Twitter JSON data. Tweets are represented as JSON objects that are exactly what was returned from the Twitter API.
- Use Cases: This tool can be used to collect data from Twitter in a way that allows for the collection of tweets, users, trends, and hydrate tweet ids. Twarc is incorporated in Social Feed Manager.
- Pros and Cons: This tool allows for downloading existing tweets using a given search query. This can include keywords, as well as hashtags. It also allows for the collection of tweets as they happen using the “filter” command. The pros and uses of the tool go on, however, there may be a learning curve if one is not familiar with using code, since it is a command line application.

### *Twitter Archiving Google Sheet (TAGS)*

- URL: <https://tags.hawksey.info/>
- Overview: A free Google Sheet template that allows users to setup and run automated collections of search results from Twitter.
- Use Cases: TAGS can be used to develop multiple archives of tweets and allows for the user to archive tweets according to specific criteria including hashtags and keywords and

to have the tweets in one, searchable archive. It also archives favorited tweets of the linked Twitter account.

- Pros and Cons: This option is free and easy to use. The website is informative and gives free video setup help as well as written instruction. It also offers free support help and a forum to ask questions. Like other open source options, this is limited by Twitter's Terms of Service. The Search API for Twitter is not a complete index, but a limited index that included between 6-9 days of tweets. Unlike other tools, it does not harvest a complete json representation of the data, but writes the response into a spreadsheet format, resulting in the loss of some metadata, such as profile information.
- Export Formats: HTML

### *Social Feed Manager*

- URL: <https://gwu-libraries.github.io/sfm-ui/>
- Overview: This is an open source software that harvests social media data and web resources from Twitter, Tumblr, Flickr, and Sina Weibo. It also allows for users to harvest web resources such as images and web pages that are linked from or embedded into the social media that is being collected.
- Use Cases: This resource allows for archivists to build collections and harvest data of the social media being used, as outlined above.
- Pros and Cons: The program allows for users to harvest data based on accounts as well as searches, filters, or sample outlines on an ongoing basis or a schedule can be specified by the administrator. The program is limited in terms of what social media accounts that an institution may want to harvest from.
- Export Formats: JSON, csv, html

### *Twitter Archive*

- URL: <https://support.twitter.com/articles/20170160>
- Overview: Twitter Archive is a self-service download tool that allows users to download captures of their Twitter profile.
- Use Cases: This tool is used to capture a specific user's information. It can be used to capture Twitter information for profiles that archivists have password access to.
- Pros and Cons: This tool is easy to use and Twitter does the retrieving of information for their users. However, it can only be used to capture one profile at a time and the archivist must have password access to the profile. It cannot be used to download certain posts and it cannot be used to automatically capture posts. It is a user initiated tool, and it does not include JSON or copies of embedded images, which remain on Twitter's server.
- Export Formats: HTML

### *Facebook Account Download*

- URL: <https://www.facebook.com/help/302796099745838>

- Overview: The Facebook Account Download is a self-service tool that allows those with Facebook accounts to download personal data from their own account.
- Use Cases: This tool would be used to copy personal data on Facebook accounts to which an archivist has access.
- Pros and Cons: This tool is an easy way to download data from a Facebook account, directly from the website. It does not, however, allow for archiving any account information that a person does not have direct password access to. It is also something that must be manually requested by the user. It cannot be used as an automatic, frequent download. This tool also does not allow for the user to pick and choose what they want to download. For example, you cannot only download posts made May 10-25, 2016 or only account data relating to the About Me section of a profile.
- Export Formats: HTML

### *WebRecorder*

- URL: <https://webrecorder.io/>
- Overview: Web Recorder is a mostly free tool that is made to create interactive recordings of websites that one browses as well as making those recordings accessible to the user, both online and offline.
- Use Cases: This tool allows for recording of individual social media accounts on the web, as the archivist moves through the web pages.
- Pros and Cons: WebRecorder not only preserves HTML and images, but according to their website, it focuses on “dynamic web content” meaning it is able to also preserve script, stylesheets, and video and audio recordings. It also now has a desktop app that allows for viewing archives offline.