# User Documentation

## 1 User Guide
- 1.1 What is SFM used for? ......................................................... 4
- 1.2 Setting up Credentials ...................................................... 6
- 1.3 Creating Collections ....................................................... 6
- 1.4 Exporting your Data ...................................................... 7

## 2 API Credentials
- 2.1 Managing credentials .................................................... 9
- 2.2 Adding Twitter Credentials ............................................. 10
- 2.3 Adding Flickr Credentials ............................................. 11
- 2.4 Adding Tumblr Credentials ............................................. 11
- 2.5 Adding Weibo Credentials ............................................. 12

## 3 Collection types
- 3.1 Twitter user timeline .................................................. 13
- 3.2 Twitter search ............................................................ 14
- 3.3 Twitter sample ............................................................ 14
- 3.4 Twitter filter .............................................................. 14
- 3.5 Flickr user ................................................................. 15
- 3.6 Tumblr blog posts ...................................................... 15
- 3.7 Weibo timeline .......................................................... 15
- 3.8 Weibo search ............................................................. 16
- 3.9 Incremental collecting ................................................... 16

## 4 Data Dictionaries for CSV/Excel Exports
- 4.1 Twitter Dictionary ....................................................... 17
- 4.2 Tumblr Dictionary ....................................................... 18
- 4.3 Flickr Dictionary ........................................................ 19
- 4.4 Weibo Dictionary ....................................................... 21

## 5 Command-line exporting/processing
- 5.1 Processing container .................................................... 23
- 5.2 SFM commandline tools .............................................. 24
- 5.3 Recipes ................................................................. 26

## 6 Releasing public datasets
- 6.1 Exporting collection data ............................................ 29
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2</td>
<td>Publishing collection data on Dataverse</td>
<td>31</td>
</tr>
<tr>
<td>6.3</td>
<td>Adding link to Dataverse dataset</td>
<td>34</td>
</tr>
<tr>
<td>7</td>
<td>Citing SFM and datasets</td>
<td>37</td>
</tr>
<tr>
<td>7.1</td>
<td>Citing SFM</td>
<td>37</td>
</tr>
<tr>
<td>7.2</td>
<td>Citing a public dataset</td>
<td>37</td>
</tr>
<tr>
<td>7.3</td>
<td>Citing your own dataset</td>
<td>38</td>
</tr>
<tr>
<td>8</td>
<td>Installation and configuration</td>
<td>39</td>
</tr>
<tr>
<td>8.1</td>
<td>Overview</td>
<td>39</td>
</tr>
<tr>
<td>8.2</td>
<td>Local installation</td>
<td>39</td>
</tr>
<tr>
<td>8.3</td>
<td>Amazon EC2 installation</td>
<td>40</td>
</tr>
<tr>
<td>8.4</td>
<td>Configuration</td>
<td>42</td>
</tr>
<tr>
<td>8.5</td>
<td>HTTPS</td>
<td>43</td>
</tr>
<tr>
<td>8.6</td>
<td>Stopping</td>
<td>43</td>
</tr>
<tr>
<td>8.7</td>
<td>Server restarts</td>
<td>43</td>
</tr>
<tr>
<td>8.8</td>
<td>Upgrading</td>
<td>43</td>
</tr>
<tr>
<td>8.9</td>
<td>Server sizing</td>
<td>44</td>
</tr>
<tr>
<td>9</td>
<td>Monitoring</td>
<td>45</td>
</tr>
<tr>
<td>9.1</td>
<td>Monitor page</td>
<td>45</td>
</tr>
<tr>
<td>9.2</td>
<td>Logs</td>
<td>45</td>
</tr>
<tr>
<td>9.3</td>
<td>RabbitMQ management console</td>
<td>46</td>
</tr>
<tr>
<td>10</td>
<td>Administration</td>
<td>47</td>
</tr>
<tr>
<td>10.1</td>
<td>Managing groups</td>
<td>47</td>
</tr>
<tr>
<td>10.2</td>
<td>Deactivating collections</td>
<td>47</td>
</tr>
<tr>
<td>10.3</td>
<td>Sharing collections</td>
<td>48</td>
</tr>
<tr>
<td>10.4</td>
<td>Deleting items</td>
<td>48</td>
</tr>
<tr>
<td>10.5</td>
<td>Moving collections</td>
<td>48</td>
</tr>
<tr>
<td>10.6</td>
<td>Allowing access to Admin Interface</td>
<td>48</td>
</tr>
<tr>
<td>11</td>
<td>Accounts and authorization</td>
<td>49</td>
</tr>
<tr>
<td>11.1</td>
<td>Accounts</td>
<td>49</td>
</tr>
<tr>
<td>11.2</td>
<td>Groups</td>
<td>49</td>
</tr>
<tr>
<td>11.3</td>
<td>Authorizations</td>
<td>49</td>
</tr>
<tr>
<td>12</td>
<td>Docker</td>
<td>51</td>
</tr>
<tr>
<td>12.1</td>
<td>Installing Docker</td>
<td>51</td>
</tr>
<tr>
<td>12.2</td>
<td>Helpful commands</td>
<td>51</td>
</tr>
<tr>
<td>12.3</td>
<td>Scaling up with Docker</td>
<td>52</td>
</tr>
<tr>
<td>12.4</td>
<td>Stopping Docker from automatically starting</td>
<td>52</td>
</tr>
<tr>
<td>13</td>
<td>Collection set / Collection portability</td>
<td>55</td>
</tr>
<tr>
<td>13.1</td>
<td>Overview</td>
<td>55</td>
</tr>
<tr>
<td>13.2</td>
<td>Preparing to move a collection set / collection</td>
<td>56</td>
</tr>
<tr>
<td>13.3</td>
<td>Loading a collection set / collection</td>
<td>56</td>
</tr>
<tr>
<td>13.4</td>
<td>Moving an entire SFM instance</td>
<td>57</td>
</tr>
<tr>
<td>14</td>
<td>Storage</td>
<td>59</td>
</tr>
<tr>
<td>14.1</td>
<td>Storage volumes</td>
<td>59</td>
</tr>
<tr>
<td>14.2</td>
<td>Volume types</td>
<td>59</td>
</tr>
<tr>
<td>14.3</td>
<td>File ownership</td>
<td>60</td>
</tr>
<tr>
<td>14.4</td>
<td>Directory structure of sfm-data</td>
<td>60</td>
</tr>
<tr>
<td>14.5</td>
<td>Space warnings</td>
<td>60</td>
</tr>
</tbody>
</table>
Social Feed Manager is open source software for libraries, archives, cultural heritage institutions and research organizations. It empowers those communities’ researchers, faculty, students, and archivists to define and create collections of data from social media platforms. Social Feed Manager will harvest from Twitter, Tumblr, Flickr, and Sina Weibo and is extensible for other platforms.

This site provides documentation for installation and usage of SFM. See the Social Feed Manager project site for full information about the project’s objectives, roadmap, and updates.
Welcome to Social Feed Manager!

Social Feed Manager (SFM) is an open-source tool designed for researchers, archivists, and curious individuals to collect social media data from Twitter, Tumblr, Flickr, or Sina Weibo. See the SFM Overview for a quick look at SFM.

If you want to learn more about what SFM can do, read What is SFM used for? This guide is for users who have access to SFM and want to learn how to collect. If you’re an administrator setting up SFM for your institution, see Admin and Technical Documentation.

To get your first collection up and running:

• **Sign up**: On the SFM homepage, click “Sign up.” Fill out the form, including a unique email. Once you sign up, you will be automatically logged in.

• **Get credentials**: You’ll need to authorize access to the social media platforms using credentials. See Setting up Credentials.

• **Create a collection set** and within it a collection, where you’ll actually collect data. See Creating Collections.

• **Add seeds**: Seeds are the criteria used to collect data. You’ll add user accounts or search criteria. See Adding Seeds.

• **Set your collections running**!

• **Export your collections** when you want to see and work with your data, or adjust settings. See Exporting your Data.

You can always come back to this user guide for help by clicking Documentation at the bottom of any SFM page and selecting User Guide.

What is SFM used for?

Types of Collections

How to use the data

Privacy and platform policy considerations
1.1 What is SFM used for?

Social Feed Manager (SFM) collects individual posts—tweets, photos, blogs—from social media sites. These posts are collected in their native, raw data format called JSON and can be exported in many formats, including spreadsheets. Users can then use this collected data for research, analysis or archiving.

Some ideas for how to use SFM:

- **Collecting from individual accounts** such as the tweets of every U.S. Senator *(Twitter user timeline)*.
- **Gathering Flickr images for analysis** or archiving the photographs from accounts donated to your organization *(Flickr user)*.
- **Researching social media use** by retrieving a sample of all tweets *(Twitter sample)*, or by filtering by specific search terms *(Twitter filter)*.
- **Capturing a major event** by collecting tweets in a specific geographic location or by following specific hashtags.
- **Collecting Tumblr posts** for preserving institutional blogs or the work of online artists *(Tumblr blog posts)*.
- **Archiving posts** from any social media platform for later research.

Note that SFM currently collects social media data from Twitter, Tumblr, Flickr, and Sina Weibo.

Here’s a sample of what a collection set looks like:
1.1.1 Types of Collections

- **Twitter user timeline**: Collect tweets from specific Twitter accounts
- **Twitter search**: Collects tweets by a user-provided search query from recent tweets
- **Twitter sample**: Collects a Twitter-provided stream of a subset of all tweets in real time.
- **Twitter filter**: Collects tweets by user-provided criteria from a stream of tweets in real time.
- **Flickr user**: Collects posts and photos from specific Flickr accounts
- **Weibo timeline**: Collects posts from the user and the user’s friends
- **Tumblr blog posts**: Collects blog posts from specific Tumblr blogs

1.1.2 How to use the data

Once you’ve collected data, there are a few ways to use it:

- You could export it into a CSV or Excel format for a basic analysis (*Exporting your Data*), or load the format into analysis software such as Stata, SPSS, or Gephi.
- You could set up an archive using the JSON files or Excel files.

1.1.3 Privacy and platform policy considerations

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.

**Ethical considerations**

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. In addition, take a look at these social media research ethical and privacy guidelines.

### 1.2 Setting up Credentials

Before you can start collecting, you need **credentials** for the social media platform that you want to use. Credentials are keys used by each platform to control the data they release to you.

You are responsible for creating your own credentials so that you can control your own collection rate and make sure that you are following the policies of each platform.

For more information about platform-specific policies, consult the documentation for each social media platform’s API.

*Adding Twitter Credentials*
*Adding Flickr Credentials*
*Adding Tumblr Credentials*
*Adding Weibo Credentials*

### 1.3 Creating Collections

**Collections** are the basic SFM containers for social media data. Each collection either gathers posts from individual accounts or gathers posts based on search criteria.

Collections are contained in **collection sets**. While collection sets sometimes only include one collection, sets can be used to organize all of the data from a single project or archive—for example, a collection set about a band might include a collection of the Twitter user timelines of each band member, a collection of the band’s Flickr, and a Twitter Filter collection of tweets that use the band’s hashtag.

Before you begin collecting, you may want to consider these **collection development guidelines**.

#### 1.3.1 Setting up Collections and Collection Sets

Because collections are housed in collection sets, you must make a collection set first.

Navigate to the Collection Sets page from the top menu, then click the **Add Collection Set** button.

Give the collection set a unique name and description. A collection set is like a folder for all collections in a project.

If you are part of a group project, you can contact your SFM administrator and set up a new group which you can share each collection set with. (This can be changed or added later on).
Once you are in a collection set, click the “Add Collection” dropdown menu and select the collection type you want to add.

Enter a unique collection name and a short description. The description is a great location to describe how you chose what to put in your collection.

Select which credential you want to use. If you need to set up new credentials, see Setting up Credentials.

### 1.3.2 Adding Seeds

**Seeds** are the criteria used by SFM to collect social media posts. Seeds may be individual social media accounts or search terms used to filter posts.

The basic process for adding seeds is the same for every collection type, except for Twitter Sample and Sina Weibo:

- Turn off the collection.
- Click Add Seed for adding one seed or Add Bulk Seeds for multiple.
- Enter either the user ids or search criteria and save.
- When you have added all seeds you want, click Turn on.

For details on each collection type, see:

- Twitter user timeline
- Twitter search
- Twitter sample
- Twitter filter
- Flickr user
- Weibo timeline
- Tumblr blog posts

### 1.4 Exporting your Data

In order to access the data in a collection, you will need to export it. You are able to download your data in several formats, including Excel (.xlsx) and Comma Separated Values (.csv), which can be loaded into a spreadsheet or data analytic software.

**To export:**

- At the top of the individual collection, click Export.
- Select the file type you want (.csv is recommended; .xlsx types will also be easily accessible).
- Select the export file size you want, based on number of posts per file. You may want to select a number of posts that will work the program that you will be loading the data into, e.g., Excel.
- Select Deduplicate if you only want one instance of every post. This will clean up your data, but will make the export take longer.
- Item start date/end date allow you to limit the export based on the date each post was created. Note that the date you enter will be in the local timezone. The date in posts may be in a different timezone, e.g., UTC. Appropriate adjustments will be made to account for this.
• Harvest start date/end date allow you to limit the export based on the harvest dates.

• When you have the settings you want, click Export. You will be redirected to the export screen. When the export is complete, the files, along with a README file describing what was included in the export and the collection, will appear for you to click on and download. You will receive an email when your export completes.

• To help understand each metadata field in the export, see Data Dictionaries for CSV/Excel Exports.
Accessing the APIs of social media platforms requires credentials for authentication (also known as API keys). Social Feed Manager supports managing those credentials. Credentials/authentication allow a user to collect data through a platform’s API. For some social media platforms (e.g., Twitter and Tumblr), limits are placed on methods and rate of collection on a per credential basis.

SFM users are responsible for creating their own new credentials so that they can control their own collection rates and can ensure that they are following each platform’s API policies.

Most API credentials have two parts: an application credential and a user credential. (Flickr is the exception – only an application credential is necessary.)

For more information about platform-specific policies, consult the documentation for each social media platform’s API.

## 2.1 Managing credentials

SFM supports two approaches to managing credentials: adding credentials and connecting credentials. Both of these options are available from the Credentials page.

### 2.1.1 Adding credentials

For this approach, a user gets the application and/or user credential from the social media platform and provides them to SFM by completing a form. More information on getting credentials is below.

### 2.1.2 Connecting credentials

*This is the easiest approach for users.*

For this approach, SFM is configured with the application credentials for the social media platform by the systems administrator. The user credentials are obtained by the user being redirected to the social media website to give permission to SFM to access her account.
SFM is configured with the application credentials in the `.env` file. If additional management is necessary, it can be performed using the Social Accounts section of the Admin interface.

### 2.2 Adding Twitter Credentials

As a user, the easiest way to set up Twitter credentials is to connect them to your personal Twitter account or another Twitter account you control. If you want more fine-tuned control, you can manually set up application-level credentials (see below). To connect Twitter credentials, first sign in to Twitter with the account you want to use. Then, on the Credentials page, click *Connect to Twitter*. Your browser will open a page from Twitter, asking you for authorization. Click *Authorize*, and your credentials will automatically connect. Once credentials are connected, you can start *Creating Collections*.

Twitter application credentials can be obtained from the Twitter API. This process requires applying for a developer account for your organization or your personal use and describing your use case for SFM. Be sure to answer all of the questions in the application. You may receive email follow-up requesting additional information before the application is approved.

Creating application credentials and manually adding Twitter credentials, rather than connecting them automatically using your Twitter account (see above), gives you greater control over your credentials and allows you to use multiple credentials.

**To obtain application credentials:**

- Sign in to Twitter.
- Follow the prompts to describe your intended use case for academic research.
- When a description for your app is requested, you may include: *This is an instance of Social Feed Manager, a social media research and archival tool, which collects data for academic researchers through an accessible user interface.*
- Enter a website such as the Social Feed Manager URL. Any website will work.
- You must provide a callback URL which is http://<SFM hostname>/accounts/twitter/login/callback/. Note that the URL should begin with `http`, not `https`, even if you are using `https`.
- Turn on Enable Callback Locking and Allow this application to be used to Sign in with Twitter.
- It is recommended to change the application permissions to read-only.
- Review and agree to the Twitter Developer Agreement.

You may need to wait several days for the account and app to be approved. One approved, it is recommended that you:

- Click on your new application.
- Navigate to the *Permissions* tab.
- Select *Read only* then *Update settings*.

You now have application-level credentials you can use in your `.env` file.

**To manually add a Twitter Credential in your SFM user account:**

- Go to the Credentials page of SFM, and click *Add Twitter Credential*.
- Fill out all fields:
  - Navigate to the *Keys and Access Tokens* tab.
– From the top half of the page, copy and paste into the matching fields in SFM: Consumer Key and Consumer Secret.
– From the bottom half of the page, copy and paste into the matching fields in SFM: Access Token and Access Token Secret.

• Click Save

2.3 Adding Flickr Credentials

• Navigate to https://www.flickr.com/services/api/keys/.
• Sign in to your Yahoo! account.
• Click Get Another Key
• Choose Apply for a Non-commercial key, which is for API users that are not charging a fee.
• Enter an Application Name like Social Feed Manager
• Enter Application Description such as: This is a social media research and archival tool, which collects data for academic researchers through an accessible user interface.
• Check both checkboxes
• Click Submit

• Navigate to the SFM Credentials page and click Add Flickr Credential
• Enter the Key and Secret in the correct fields and save.

2.4 Adding Tumblr Credentials

• Navigate to https://www.tumblr.com/oauth/apps/.
• Sign in to Tumblr.
• Click Register Application
• Enter an Application Name like Social Feed Manager
• Enter a website such as the SFM url
• Enter Application Description such as: This is a social media research and archival tool, which collects data for academic researchers through an accessible user interface.
• Enter Administrative contact email. You should use your own email.
• Enter default callback url, the same url used for the website.
• Click Register

• Navigate to the SFM Credentials page and click Add Tumblr Credential
• Enter the OAuth Consumer Key in the API key field and save.
2.5 Adding Weibo Credentials

For instructions on obtaining Weibo credentials, see this guide.

To use the connecting credentials approach for Weibo, the redirect URL must match the application’s actual URL and use port 80.
Collection types

Each collection type connects to one of a social media platform’s APIs, or methods for retrieving data. Understanding what each collection type provides is important to ensure you collect what you need and are aware of any limitations. Reading the social media platform’s documentation provides further important details.

Collection types

- **Twitter user timeline**: Collect tweets from specific Twitter accounts
- **Twitter search**: Collects tweets by a user-provided search query from recent tweets
- **Twitter sample**: Collects a Twitter provided stream of a subset of all tweets in real time.
- **Twitter filter**: Collects tweets by user-provided criteria from a stream of tweets in real time.
- **Flickr user**: Collects posts and photos from specific Flickr accounts
- **Weibo timeline**: Collects posts from the user and the user’s friends
- **Weibo search**: Collects recent weibo posts by a user-provided search query
- **Tumblr blog posts**: Collects blog posts from specific Tumblr blogs

### 3.1 Twitter user timeline

Twitter user timeline collections collect the 3,200 most recent tweets from each of a list of Twitter accounts using Twitter’s user_timeline API.

**Seeds** for Twitter user timelines are individual Twitter accounts.

To identify a user timeline, you can provide a screen name (the string after @, like NASA for @NASA) or Twitter user ID (a numeric string which never changes, like 11348282 for @NASA). If you provide one identifier, the other will be looked up and displayed in SFM the first time the harvester runs. The user may change the screen name over time, and the seed will be updated accordingly.

The harvest schedule should depend on how prolific the Twitter users are. In general, the more frequent the tweeter, the more frequent you’ll want to schedule harvests.
SFM will notify you when incorrect or private user timeline seeds are requested; all other valid seeds will be collected. See *Incremental collecting* to decide whether or not to collect incrementally.

### 3.2 Twitter search

Twitter searches collect tweets from the last 7-9 days that match search queries, similar to a regular search done on Twitter, using the [Twitter Search API](https://developer.twitter.com/en/docs/search/api). This is *not* a complete search of all tweets; results are limited both by time and arbitrary relevance (determined by Twitter).

Search queries must follow standard search term formulation; permitted queries are listed in the documentation for the Twitter Search API, or you can construct a query using the Twitter Advanced Search query builder.

Broad Twitter searches may take longer to complete – possibly days – due to Twitter’s rate limits and the amount of data available from the Search API. In choosing a schedule, make sure that there is enough time between searches. (If there is not enough time between searches, later harvests will be skipped until earlier harvests complete.) In some cases, you may only want to run the search once and then turn off the collection. See *Incremental collecting* to decide whether or not to collect incrementally.

### 3.3 Twitter sample

Twitter samples are a random collection of approximately 0.5–1% of public tweets, using the [Twitter sample stream](https://developer.twitter.com/en/docs/streaming-api/guides/sample), useful for capturing a sample of what people are talking about on Twitter. The Twitter sample stream returns approximately 0.5-1% of public tweets, which is approximately 3GB a day (compressed).

Unlike other Twitter collections, there are no seeds for a Twitter sample.

When on, the sample returns data every 30 minutes.

Only one sample or *Twitter filter* can be run at a time per credential.

### 3.4 Twitter filter

Twitter Filter collections harvest a live selection of public tweets from criteria matching keywords, locations, languages, or users, based on the [Twitter filter streaming API](https://developer.twitter.com/en/docs/streaming-api/guides/filters). Because tweets are collected live, tweets from the past are not included. (Use a Twitter search collection to find tweets from the recent past.)

There are four different filter queries supported by SFM: track, follow, location, and language.

**Track** collects tweets based on a keyword search. A space between words is treated as ‘AND’ and a comma is treated as ‘OR’. Note that exact phrase matching is not supported. See the *track parameter documentation* for more information.

- **Note:** When entering a comma-separated list of search terms for the track or follow parameters, make sure to use the standard `,` character. When typing in certain languages that use a non-Roman alphabet, a different character is generated for commas. For example, when typing in languages such as Arabic, Farsi, Urdu, etc., typing a comma generates the `،` character. To avoid errors, the Track parameter should use the Roman `,` character; for example: `，`.

**Follow** collects tweets that are posted by or about a user (not including mentions) from a comma separated list of user IDs (the numeric identifier for a user account). Tweets collected will include those made by the user, retweeting the user, or replying to the user. See the *follow parameter documentation* for more information.
• Note: The Twitter website does not provide a way to look up the user ID for a user account. You can use https://tweeterid.com for this purpose.

**Location** collects tweets that were geolocated within specific parameters, based on a bounding box made using the southwest and northeast corner coordinates. See the location parameter documentation for more information.

**Language** collects tweets that Twitter detected as being written in the specified languages. For example, specifying *en,es* will only collect Tweets detected to be in the English or Spanish languages. See the language parameter documentation for more information.

Twitter will return a limited number of tweets, so filters that return many results will not return all available tweets. Therefore, more narrow filters will usually return more complete results.

Only one filter or Twitter sample can be run at a time per credential.

SFM captures the filter stream in 30 minute chunks and then momentarily stops. Between rate limiting and these momentary stops, you should never assume that you are getting every tweet.

There is only one seed in a filter collection. Twitter filter collection are either turned on or off (there is no schedule).

### 3.5 Flickr user

Flickr User Timeline collections gather metadata about public photos by a specific Flickr user, and, optionally, copies of the photos at specified sizes.

Each Flickr user collection can have multiple seeds, where each seed is a Flickr user. To identify a user, you can provide a either a username or an NSID. If you provide one, the other will be looked up and displayed in the SFM UI during the first harvest. The NSID is a unique identifier and does not change; usernames may be changed but are unique.

Usernames can be difficult to find, so to ensure that you have the correct account, use this tool to find the NSID from the account URL (i.e., the URL when viewing the account on the Flickr website).

Depending on the image sizes you select, the actual photo files will be collected as well. Be very careful in selecting the original file size, as this may require a significant amount of storage. Also note that some Flickr users may have a large number of public photos, which may require a significant amount of storage. It is advisable to check the Flickr website to determine the number of photos in each Flickr user’s public photo stream before harvesting.

For each user, the user’s information will be collected using Flickr’s people.getInfo API and the list of her public photos will be retrieved from people.getPublicPhotos. Information on each photo will be collected with photos.getInfo.

See Incremental collecting to decide whether or not to collect incrementally.

### 3.6 Tumblr blog posts

Tumblr Blog Post collections harvest posts by specified Tumblr blogs using the Tumblr Posts API.

**Seeds** are individual blogs for these collections. Blogs can be specified with or without the .tumblr.com extension.

See Incremental collecting to decide whether or not to collect incrementally.

### 3.7 Weibo timeline

Weibo Timeline collections harvest weibos (microblogs) by the user and friends of the user whose credentials are provided using the Weibo friends_timeline API.
Note that because collection is determined by the user whose credentials are provided, there are no seeds for a Weibo timeline collection. To change what is being collected, change the user’s friends from the Weibo website or app.

### 3.8 Weibo search

Collects recent weibos that match a search query using the Weibo search_topics API. The Weibo API does not return a complete search of all Weibo posts. It only returns the most recent 200 posts matching a single keyword when found between pairs of `#` in Weibo posts (for example: `#keyword#` or `##`)

The incremental option will attempt to only count weibo posts that haven’t been harvested before, maintaining a count of non-duplicate weibo posts. Because the Weibo search API does not accept `since_id` or `max_id` parameters, filtering out already-harvested weibos from the search count is accomplished within SFM.

When the incremental option is not selected, the search will be performed again, and there will most likely be duplicates in the count.

### 3.9 Incremental collecting

The incremental option is the default and will collect tweets or posts that have been published since the last harvest. When the incremental option is not selected, the maximum number of tweets or posts will be harvested each time the harvest runs. If a non-incremental harvest is performed multiple times, there will most likely be duplicates. However, with these duplicates, you may be able to track changes across time in a user’s timeline, such as changes in retweet and like counts, deletion of tweets, and follower counts.
Social Feed Manager captures a variety of data from each platform. These data dictionaries give explanations for each selected and processed field in exports.

Note that these are subsets of the data that are collected for each post. The full data is available for export by selecting “Full JSON” as the export format or by exporting from the commandline. See Command-line exporting/processing.

- Twitter Dictionary
- Tumblr Dictionary
- Flickr Dictionary
- Weibo Dictionary

### 4.1 Twitter Dictionary

For more info about source tweet data, see the Twitter API documentation, including Tweet data dictionaries.

Documentation about older archived tweets is archived by the Wayback Machine for the Twitter API, Tweets, and Entities.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Twitter identifier for the tweet.</td>
</tr>
<tr>
<td>tweet_url</td>
<td>URL of the tweet on Twitter’s website. If the tweet is a retweet, the URL will be redirected to the original tweet.</td>
</tr>
<tr>
<td>created_at</td>
<td>Date and time the tweet was created, in Twitter’s default format.</td>
</tr>
<tr>
<td>parsed_created_at</td>
<td>Date and time the tweet was created, in ISO 8601 format and UTC time zone.</td>
</tr>
<tr>
<td>user_screen_name</td>
<td>The unique screen name of the account that authored the tweet, at the time the tweet was posted. Scree</td>
</tr>
<tr>
<td>text</td>
<td>The text of the tweet. Newline characters are replaced with a space.</td>
</tr>
<tr>
<td>tweet_type</td>
<td>original, reply, quote, or retweet</td>
</tr>
<tr>
<td>coordinates</td>
<td>The geographic coordinates of the tweet. This is only enabled if geotagging is enabled on the account.</td>
</tr>
<tr>
<td>hashtags</td>
<td>Hashtags from the tweet text, as a comma-separated list. Hashtags are generally displayed with a # symbol.</td>
</tr>
<tr>
<td>media</td>
<td>URLs of media objects (photos, videos, GIFs) that are attached to the tweet.</td>
</tr>
<tr>
<td>Field</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>urls</td>
<td>URLs entered by user as part of tweet. Note that URL may be a shortened URL, e.g. from bit.ly.</td>
</tr>
<tr>
<td>favorite_count</td>
<td>Number of times this tweet had been favorited/liked by other users at the time the tweet was collected.</td>
</tr>
<tr>
<td>in_reply_to_screen_name</td>
<td>If tweet is a reply, the screen name of the author of the tweet that is being replied to.</td>
</tr>
<tr>
<td>in_reply_to_status_id</td>
<td>If tweet is a reply, the Twitter identifier of the tweet that is being replied to.</td>
</tr>
<tr>
<td>in_reply_to_user_id</td>
<td>If tweet is a reply, the Twitter identifier of the author of the tweet that is being replied to.</td>
</tr>
<tr>
<td>lang</td>
<td>Language of the tweet text, as determined by Twitter.</td>
</tr>
<tr>
<td>place</td>
<td>The user or application-provided geographic location from which a tweet was posted.</td>
</tr>
<tr>
<td>possibly_sensitive</td>
<td>Indicates that URL contained in the tweet may reference sensitive content.</td>
</tr>
<tr>
<td>retweet_count</td>
<td>Number of times the tweet had been retweeted at the time the tweet was collected.</td>
</tr>
<tr>
<td>retweet_or_quote_id</td>
<td>If tweet is a retweet or quote tweet, the Twitter identifier of the source tweet.</td>
</tr>
<tr>
<td>retweet_or_quote_screen_name</td>
<td>If tweet is a retweet or quote tweet, the screen name of the author of the source tweet.</td>
</tr>
<tr>
<td>retweet_or_quote_user_id</td>
<td>If tweet is a retweet or quote tweet, the Twitter identifier of the author or the source tweet.</td>
</tr>
<tr>
<td>source</td>
<td>The application from which the tweet was posted.</td>
</tr>
<tr>
<td>user_id</td>
<td>Twitter identifier for the author of the tweet.</td>
</tr>
<tr>
<td>user_created_at</td>
<td>Date and time the tweet was created, in Twitter’s default format.</td>
</tr>
<tr>
<td>user_default_profile_image</td>
<td>URL of the user’s profile image.</td>
</tr>
<tr>
<td>user_description</td>
<td>The user-provided account description. Newline characters are replaced with a space.</td>
</tr>
<tr>
<td>user_favourites_count</td>
<td>Number of tweets that have been favorited/liked by the user.</td>
</tr>
<tr>
<td>user_followers_count</td>
<td>Number of followers this account had at the time the tweet was collected.</td>
</tr>
<tr>
<td>user_friends_count</td>
<td>Number of users this account was following at the time the tweet was collected.</td>
</tr>
<tr>
<td>user_listed_count</td>
<td>Number of public lists that this user is a member of.</td>
</tr>
<tr>
<td>user_location</td>
<td>The user’s self-described location. Not necessarily an actual place.</td>
</tr>
<tr>
<td>user_name</td>
<td>The user’s self-provided name.</td>
</tr>
<tr>
<td>user_statuses_count</td>
<td>Number of tweets that the user has posted.</td>
</tr>
<tr>
<td>user_timezone</td>
<td>The user-provided time zone. Currently deprecated.</td>
</tr>
<tr>
<td>user_urls</td>
<td>URLs entered by user as part of user’s description.</td>
</tr>
<tr>
<td>user_verified</td>
<td>Indicates that the user’s account is verified.</td>
</tr>
</tbody>
</table>

### 4.2 Tumblr Dictionary

For more info about source tweet data, see the [Tumblr API documentation](#), particularly Posts.

Documentation about older archived posts is archived by the Wayback Machine for the original Tumblr API and the newer Tumblr API.
### 4.3 Flickr Dictionary

For more info about source tweet data, see the Flickr API documentation, particularly People and Photos.

Documentation about older archived posts is archived by the Wayback Machine here.
<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>photo_id</td>
<td>Unique Flickr identifier of the photo.</td>
<td>11211844604</td>
</tr>
<tr>
<td>date_posted</td>
<td>Date and time that the post was uploaded to Flickr, in ISO 8601 format and UTC time zone.</td>
<td>2013-12-04 21:39:40+00:00</td>
</tr>
<tr>
<td>date_taken</td>
<td>Date and time that media was captured, either extracted from EXIF or from the date posted, in mm/dd/yyyy hh:mm format.</td>
<td>6/7/2014 13:35</td>
</tr>
</tbody>
</table>
| license      | Licensing allowed for media, given as a numeral according to the following key:  
|              | • 0 = All Rights Reserved  
|              | • 1 = Attribution-NonCommercial-Sharealike License  
|              | • 2 = Attribution-NonCommercial License  
|              | • 3 = Attribution-NonCommercial NoDerivs License  
|              | • 4 = Attribution License  
|              | • 5 = Attribution-ShareAlike License  
|              | • 6 = Attribution-NoDerivs License  
|              | • 7 = No known copyright restrictions  
|              | • 8 = United States Government work  
|              | • More information at creativecommons.org/licenses                              | 4 (Attribution license)                      |
| safety_level | Appropriateness of post, given as a numeral according to the following key:  
|              | • 0 = Safe - Content suitable for everyone  
|              | • 1 = Moderate - Approximately PG-13 content  
|              | • 2 = Restricted - Approximately R rated content                                | 0 (Safe level)                               |
| original_format | File format of uploaded media.                                      | jpg                                          |
| owner_nsid   | Unique Flickr identifier of the owner account.                              | 28399705@N04                                 |
| owner_username | Unique plaintext username of the owner account.                           | GW Museum and Textile Museum                 |
| title        | Title of the post.                                                        | Original Museum entrance                     |
| description  | Short description of the post.                                             | Historic photo courtesy of The Textile Museum Archives. |
| media        | Media type of the post.                                                   | photo                                        |
| photopage    | Location url of the post.                                                 | https://www.flickr.com/photos/textilemuseum/11211844604/ |
4.4 Weibo Dictionary

For more info about source tweet data, see the Sina Weibo API friends_timeline documentation.

Documentation about older archived tweets is archived by the Wayback Machine here.

Note that for privacy purposes, Weibo dictionary examples are not consistent.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>created_at</td>
<td>Date and time the tweet was created, in ISO 8601 format and UTC time zone.</td>
<td>2016-12-21T19:30:03+00:00</td>
</tr>
<tr>
<td>weibo_id</td>
<td>Sina Weibo identifier for the tweet.</td>
<td>4060309792585658</td>
</tr>
<tr>
<td>screen_name</td>
<td>The unique screen name of the account that authored the weibo, at the time the weibo was posted.</td>
<td></td>
</tr>
<tr>
<td>followers_count</td>
<td>Number of followers this account had at the time the weibo was harvested.</td>
<td>3655329</td>
</tr>
<tr>
<td>friends_count</td>
<td>Number of users this account was following at the time the weibo was harvested.</td>
<td>2691</td>
</tr>
<tr>
<td>reposts_count</td>
<td>Number of times this weibo had been reposted at the time the weibo was harvested.</td>
<td>68</td>
</tr>
<tr>
<td>topics</td>
<td>Topics (similar to hashtags) from the weibo text as a comma-separated list.</td>
<td></td>
</tr>
<tr>
<td>in_reply_to_screen_name</td>
<td>If the weibo is a reply, the screen name of the original weibo’s author. (This is not yet supported by Sina Weibo.)</td>
<td></td>
</tr>
<tr>
<td>weibo_url</td>
<td>URL of the weibo. If the tweet is a retweet made</td>
<td><a href="http://m.weibo.cn/1618051664/4060300716095462">http://m.weibo.cn/1618051664/4060300716095462</a></td>
</tr>
<tr>
<td>text</td>
<td>The text of the weibo.</td>
<td></td>
</tr>
<tr>
<td>url1</td>
<td>First URL in text of weibo, as shortened by Sina Weibo.</td>
<td><a href="http://t.cn/RM2xyx6">http://t.cn/RM2xyx6</a></td>
</tr>
<tr>
<td>url2</td>
<td>Second URL in text of weibo, as shortened by Sina Weibo.</td>
<td><a href="http://t.cn/Rc52gDY">http://t.cn/Rc52gDY</a></td>
</tr>
<tr>
<td>retweeted_text</td>
<td>Text of original weibo when the collected weibo is a repost.</td>
<td></td>
</tr>
<tr>
<td>retweeted_url1</td>
<td>First URL in text of original weibo, as shortened by Sina Weibo.</td>
<td><a href="http://t.cn/RVR4cAQ">http://t.cn/RVR4cAQ</a></td>
</tr>
<tr>
<td>retweeted_url2</td>
<td>Second URL in text of original weibo, as shortened by Sina Weibo.</td>
<td><a href="http://t.cn/RMAJISP">http://t.cn/RMAJISP</a></td>
</tr>
</tbody>
</table>
CHAPTER 5

Command-line exporting/processing

While social media data can be exported from the SFM UI, in some cases you may want to export from the command-line. These cases include:

• Exporting very large datasets. (Export via the UI is performed serially; export via the commandline can be performed in parallel, which may be much faster.)
• Performing more advanced filtering or transformation that is not supported by the UI export.
• Integrating with a processing/analysis pipeline.

To support export and processing from the commandline, SFM provides a processing container. A processing container is a Linux shell environment with access to the SFM’s data and preloaded with a set of useful tools.

Using a processing container requires familiarity with the Linux shell and shell access to the SFM server. If you are interested in using a processing container, please contact your SFM administrator for help.

When exporting/processing data, remember that harvested social media content are stored in /sfm-data. /sfm-processing is provided to store your exports, processed data, or scripts. Depending on how it is configured, you may have access to /sfm-processing from your local filesystem. See Storage.

5.1 Processing container

To bootstrap export/processing, a processing image is provided. A container instantiated from this image is Ubuntu 14.04 and pre-installed with the warc iterator tools, find_warcs.py, and some other useful tools. (Warc iterators and find_warcs.py are described below.) It will also have read-only access to the data from /sfm-data and read/write access to /sfm-processing.

The other tools available in a processing container are:

• jq for JSON processing.
• twarc for access to the Twarc utils.
• JWAT Tools for processing WARCs.
• warctools for processing WARCs.
• **parallel** for parallelizing processing.
• **csvkit** for processing CSVs.
• **gron** for grepping JSON.

To instantiate a processing container, from the directory that contains your `docker-compose.yml` file:

```bash
docker-compose run --rm processing /bin/bash
```

You will then be provided with a bash shell inside the container from which you can execute commands:

```bash
root@0ac9caaf7e72:/sfm-processing#
```

```bash
find_warcs.py 4f4d1 | xargs twitter_rest_warc_iter.py | python /opt/twarc/utils/wordcloud.py
```

Note that once you exit the processing container, the container will be automatically removed. However, if you have saved all of your scripts and output files to `/sfm-processing`, they will be available when you create a new processing container.

### 5.2 SFM commandline tools

#### 5.2.1 Warc iterators

SFM stores harvested social media data in WARC files. A warc iterator tool provides an iterator to the social media data contained in WARC files. When used from the commandline, it writes out the social media items one at a time to standard out. (Think of this as `cat`-ing a line-oriented JSON file. It is also equivalent to the output of Twarc.)

Each social media type has a separate warc iterator tool. For example, `twitter_rest_warc_iter.py` extracts tweets recorded from the Twitter REST API. For example:

```bash
root@0ac9caaf7e72:/sfm-data#
twitter_rest_warc_iter.py
```

```bash
usage: twitter_rest_warc_iter.py [-h] [--pretty] [--dedupe]
[--print-item-type] 
filepaths [filepaths ...]
```

Here is a list of the warc iterators:

- `twitter_rest_warc_iter.py`: Tweets recorded from Twitter REST API.
- `twitter_stream_warc_iter.py`: Tweets recorded from Twitter Streaming API.
- `flickr_photo_warc_iter.py`: Flickr photos
- `weibo_warc_iter.py`: Weibos
- `tumblr_warc_iter.py`: Tumblr posts

Warc iterator tools can also be used as a library.

#### 5.2.2 Find Warcs

`find_warcs.py` helps put together a list of WARC files to be processed by other tools, e.g., warc iterator tools. (It gets the list of WARC files by querying the SFM API.)

Here is arguments it accepts:
For example, to get a list of the WARC files in a particular collection, provide some part of the collection id:

```
root@0ac9caaf7e72:/sfm-data# find_warcs.py 4f4d1
/sfm-data/collection_set/b06d164c632d405294d3c17584f03278/
   → 4f4d1a6677f34ad39bb8d886e22de3b/2016/05/04/14/515dab00c05740f487e095773c8a1b-
   → 2016050413638715-0000-47-88e5bc8a36a5-8000.warc.gz
```

(In this case there is only one WARC file. If there was more than one, it would be space separated. Use `--newline` to separate with a newline instead.)

The collection id can be found from the SFM UI.

Note that if you are running `find_warcs.py` from outside a Docker environment, you will need to supply `--api-base-url`.

### 5.2.3 Sync scripts

Sync scripts will extract Twitter data from WARC files for a collection and write tweets to line-oriented JSON files and tweet ids to text files. It is called a “sync script” because it will skip WARCs that have already been processed.

Sync scripts are parallelized, allowing for faster processing.

There are sync scripts for Twitter REST collections (`twitter_rest_sync.sh`) and Twitter stream collections (`twitter_stream_sync.sh`). Usage is `./<script> <collection id> <destination directory> <# of threads>`. For example:

```
cd /opt/processing
mkdir /sfm-processing/test
./twitter_rest_sync.sh e76b140351574015a6aac8999b06dca7 /sfm-processing/test 2
```

### 5.2.4 READMEs

The `exportreadme` management command will output a README file that can be used as part of the documentation for a dataset. The README contains information on the collection, including the complete change log. Here is an example of creating a README:

```
docker-compose exec ui /bin/bash -c "/opt/sfm-ui/sfm/manage.py exportreadme 4f4d1 > /%
   → sfm-processing/README.txt"
```

For examples, see the README files in [this open dataset](https://example.com).

Note that this is a management command; thus, it is executed differently than the commandline tools described above.

---

### 5.2. SFM commandline tools

---

---
5.3 Recipes

5.3.1 Extracting URLs

The “Extracting URLs from #PulseNightclub for seeding web archiving” blog post provides some useful guidance on extracting URLs from tweets, including unshortening and sorting/counting.

5.3.2 Exporting to line-oriented JSON files

This recipe is for exporting social media data from WARC files to line-oriented JSON files. There will be one JSON file for each WARC. This may be useful for some processing or for loading into some analytic tools.

This recipe uses parallel for parallelizing the export.

Create a list of WARC files:

```
find_warcs.py --newline 7c37157 > source.lst
```

Replace 7c37157 with the first few characters of the collection id that you want to export. The collection id is available on the collection detail page in SFM UI.

Create a list of JSON destination files:

```
cat source.lst | xargs basename -a | sed 's/.warc.gz/.json/' > dest.lst
```

This command puts all of the JSON files in the same directory, using the filename of the WARC file with a .json file extension.

If you want to maintain the directory structure, but use a different root directory:

```
cat source.lst | sed 's/sfm-data\//collection_set/sfm-processing\//export/' | sed 's/.warc.gz/.json/'
```

Replace sfm-processing/export with the root directory that you want to use.

Perform the export:

```
parallel -a source.lst -a dest.lst --xapply "twitter_stream_warc_iter.py {1} > {2}"
```

Replace twitter_stream_warc_iter.py with the name of the warc iterator for the type of social media data that you are exporting.

You can also perform a filter on export using jq. For example, this only exports tweets in Spanish:

```
parallel -a source.lst -a dest.lst --xapply "twitter_stream_warc_iter.py {1} | jq -c \"select(.lang == \"es\")\" > {2}"
```

And to save space, the JSON files can be gzip compressed:

```
parallel -a source.lst -a dest.lst --xapply "twitter_stream_warc_iter.py {1} | gzip > {2}"
```

You might also want to change the file extension of the destination file to “.json.gz” by adjusting the command used to create the list of JSON destination files. To access the tweets in a gzipped JSON file, use:

```
gzip -c <filepath>
```
5.3.3 Counting posts

`wc -l` can be used to count posts. To count the number of tweets in a collection:

```
find_warcs.py 7c37157 | xargs twitter_stream_warc_iter.py | wc -l
```

To count the posts from line-oriented JSON files created as described above:

```
cat dest.lst | xargs wc -l
```

`wc -l gotcha`: When doing a lot of counting, `wc -l` will output a partial total and then reset the count. The partial totals must be added together to get the grand total. For example:

```
[Some lines skipped ...
  1490 ./964be41e1714492bbe8ec5793e05ec86-20160725070757217-00000-7932-62ebe35d576c-→8002.json
  4514 ./5f78a79c6382476889dled4734d6105a-20160722202703869-00000-5110-62ebe35d576c-→8002.json
  52043 ./417cf950a00d44408458c93f08f0690e-20160910032351524-00000-1775-c4aea5d70c14-→8000.json
54392684 total
[Some lines skipped ...
  34778 ./30bc1c34880d404aa3254f82dd387514-20160806132811173-00000-21585-→62ebe35d576c-8000.json
  30588 ./964be41e1714492bbe8ec5793e05ec86-20160727030754726-00000-10044-→62ebe35d576c-8002.json
21573971 total
```

5.3.4 Using jq to process JSON

For tips on using `jq` with JSON from Twitter and other sources, see:

- Getting Started Working with Twitter Data Using jq
- Recipes for processing Twitter data with jq
- Reshaping JSON with jq
CHAPTER 6

Releasing public datasets

Many social media platforms place limitations on sharing of data collected from their APIs. One common approach for sharing data, in particular for Twitter, is to only share the identifiers of the social media items. Someone can then recreate the dataset by retrieving the items from the API based on the identifiers. For Twitter, the process of extracting tweet ids is often called “dehydrating” and retrieving the full tweet is called “hydrating.”

Note that retrieving the entire original dataset may not be possible, as the social media platform may opt not to provide social media items that have been deleted or are no longer public.

This example shows the steps for releasing the Women’s March dataset to Dataverse. The Women’s March dataset was created by GWU and published on the Harvard Dataverse. These instructions can be adapted for publishing your own collections to the dataset repository of your choice.

Note that the Women’s March dataset is a single (SFM) collection. For an example of publishing multiple collections to a single dataset, see the 2016 United States Presidential Election dataset.

6.1 Exporting collection data

1. Access the server where your target collection is located and instantiate a processing container. (See Command-line exporting/processing):

```
ssh sfmsserver.org
cd /opt/sfm
docker-compose run --rm processing /bin/bash
```

Replace sfmsserver.org with the address of the SFM server that you want export data from.

2. Find a list of WARC files where the data of your target collection are stored, and create a list of WARC files (source.lst) and a list of destination text files. (dest.lst):

```
find_warcs.py 0110497 | tr ' ' '
' > source.lst
cat source.lst | xargs basename -a | sed 's/warc.gz/./txt/' > dest.lst
```
Replace 0110497 with the first few characters of the collection id that you want to export. The collection id is available on the collection detail page in SFM UI. (See the picture below.)

3. Write the tweet ids to the destination text files:

```
time parallel -j 3 -a source.lst -a dest.lst --xapply "twitter_stream_warc_iter.py {1} | jq -r ".id_str" > {2}"
```

This command executes a Twitter Stream WARC iterator to extract the tweets from the WARC files and jq to extract the tweet ids. This shows using `twitter_stream_warc_iter.py` for a Twitter stream collection. For a Twitter REST collection, use `twitter_rest_warc_iter.py`.

Parallel is used to perform this process in parallel (using multiple processors), using WARC files from `source.lst` and text files from `dest.lst`. `-j 3` limits parallel to 3 processors. Make sure to select an appropriate number for your server.

An alternative to steps 1 and 2 is to use a sync script to write tweet id text files and tweet JSON files in one step. (See Command-line exporting/processing)

4. Combine multiple files into large files:

The previous command creates a single text file containing tweet ids for each WARC file. To combine the tweets into a single file:

```
cat *.txt > womensmarch.txt
```

- Recommendation: If there are a large number of tweet ids in a file, split into multiple, smaller files. (We limit to 50 million tweet ids per file.)

5. Create a README file that contains information on each collection (management command for sfm ui):

Exit from the processing container, and connect to the UI container and execute the exportreadme management command to create a README file for the dataset:
6. Copy the files from the server to your local hard drive:

Exit from the SFM server with `exit` command, move to a location in your local hard drive where you want to store the data, and run the command below:

```
exit
scp -p username@sfmserver.org:/sfm-processing/womensmarch*.txt .
```

Replace `username` and `sfmsserver.org` with your user ID and the address of the SFM server, respectively.

### 6.2 Publishing collection data on Dataverse

For this example, we will be adding the collection to the GW Libraries Dataverse on the Harvard Dataverse instance.

1. Go to the GW Libraries Dataverse and log in.
   
   • Note: You should be a Curator for the dataverse to be able to upload data.

2. Open the New Dataset page:

   Click ‘Add Data > New Dataset’.
3. Fill the metadata with proper information (title, author, contact, description, subject, keyword):
   Make sure you input the right number of tweets collected and appropriate dates in the description.
4. Upload the files (both data and README files) and save the dataset:

- Note: The dataset will be saved as a draft.
5. Publish the dataset:
   Go to the page of the draft that was just saved, and click ‘Publish’ button.

6.3 Adding link to Dataverse dataset

Once you have published your collection data on Dataverse, you can add to it from SFM. This will allow other SFM users to find the public version of your collection.

1. Go to the collection page for your collection in SFM and click Edit.
2. Add the Dataverse link to in the “Public Link” field and click Save.
CHAPTER 7

Citing SFM and datasets

7.1 Citing SFM

The recommended citation for Social Feed Manager (i.e., the software) is:


For more guidance on citing SFM, see SFM in Zenodo.

7.2 Citing a public dataset

Some SFM collections have been released as public datasets, usually by depositing them in a data repository. (See Releasing public datasets).

Usually the public version will provide guidance on citing. For example, the 2016 United States Presidential Election collection is deposited in Harvard’s Dataverse, which offers the following assistance on citing:


Within SFM, a link may be provided to the public version of a dataset.
7.3 Citing your own dataset

To make your dataset citable and reusable by others, you are encouraged to release it as public dataset. (See *Releasing public datasets*). You are also encouraged to cite SFM within your dataset release and your publication.
CHAPTER 8

Installation and configuration

8.1 Overview

The supported approach for deploying SFM is Docker containers. For more information on Docker, see Docker.

Each SFM service will provide images for the containers needed to run the service (in the form of Dockerfile s). These images will be published to Docker Hub. GWU created images will be part of the GWUL organization and be prefixed with sfm-.

sfm-docker provides the necessary docker-compose.yml files to compose the services into a complete instance of SFM.

The following will describe how to setup an instance of SFM that uses the latest release (and is suitable for a production deployment.) See the development documentation for other SFM configurations.

SFM can be deployed without Docker. The various Dockerfile s should provide reasonable guidance on how to accomplish this.

8.2 Local installation

Installing locally requires Docker and Docker-Compose. See Installing Docker.

1. Either git clone the sfm-docker repository and copy the example configuration files:

```
git clone https://github.com/gwu-libraries/sfm-docker.git
cd sfm-docker
# Replace 2.3.0 with the correct version.
git checkout 2.3.0
cp example.prod.docker-compose.yml docker-compose.yml
cp example.env .env
```

or just download example.prod.docker-compose.yml and example.env (replacing 2.3.0 with the correct version):
curl -L https://raw.githubusercontent.com/gwu-libraries/sfm-docker/2.3.0/example.prod.docker-compose.yml > docker-compose.yml

curl -L https://raw.githubusercontent.com/gwu-libraries/sfm-docker/2.3.0/example.env > .env

2. Update configuration in .env as described in Configuration.

3. Download containers and start SFM:

   
   docker-compose up -d

   

It may take several minutes for the images to be downloaded and the containers to start. These images are large (roughly 12GB) so make sure you have enough disk space and a high-speed connection is recommended.

4. It is also recommended that you scale up the Twitter REST Harvester container:

   
   docker-compose scale twitterrestharvester=2 twitterpriorityrestharvester=2

   

Notes:

- The first time you bring up the containers, their images will be pulled from Docker Hub. This will take several minutes.

- For instructions on how to make configuration changes after the containers have been brought up, see Configuration.

- To learn more about scaling, see Scaling up with Docker.

- For suggestions on sizing your SFM server, see Server sizing.

- For help with other Docker commands (e.g., to stop SFM) see Helpful commands.

8.3 Amazon EC2 installation

To launch an Amazon EC2 instance running SFM, follow the normal procedure for launching an instance. In Step 3: Configure Instance Details, under Advanced Details paste the following in user details and modify as appropriate as described in Configuration. Also, in the curl statements change master to the correct version, e.g., 2.3.0:

   
   #cloud-config
   repo_update: true
   repo_upgrade: all

   packages:
   - python-pip

   runcmd:
   - curl -sSL https://get.docker.com/ | sh
   - usermod -aG docker ubuntu
   - pip install -U docker-compose
   - mkdir /sfm-data
   - mkdir /sfm-processing
   - cd /home/ubuntu
   # This brings up the latest production release. To bring up master, remove prod.
   - curl -L https://raw.githubusercontent.com/gwu-libraries/sfm-docker/2.3.0/example.prod.docker-compose.yml > docker-compose.yml
   - curl -L https://raw.githubusercontent.com/gwu-libraries/sfm-docker/2.3.0/example.env > .env

(continues on next page)
# Set config below by uncommenting.
# Don't forget to escape $ as \$.

## COMMON CONFIGURATION
- echo TZ=America/New_York >> .env

## VOLUME CONFIGURATION
Don't change these.
- echo DATA_VOLUME=/sfm-data:/sfm-data >> .env
- echo PROCESSING_VOLUME=/sfm-processing:/sfm-processing >> .env

## SFM UI CONFIGURATION
Don't change this.
- echo SFM_PORT=80 >> .env

# Provide your institution name display on sfm-ui footer
- echo SFM_INSTITUTION_NAME=yourinstitution >> .env
# Provide your institution link
- echo SFM_INSTITUTION_LINK=http://library.yourinstitution.edu >> .env
# Set to True to enable the cookie consent popup
- echo SFM_ENABLE_COOKIE_CONSENT=False >> .env
# Provide the text you would like to appear on the cookie popup
- echo SFM_COOKIE_CONSENT_HTML=<b>Do you like cookies?</b> &amp; We use cookies, to ensure you get the best experience on our website. &lt;a href="https://cookiesandyou.com/" target="_blank">Learn more</a> >> .env
# Provide the wording you would like to appear on the cookie button
- echo SFM_COOKIE_CONSENT_BUTTON_TEXT=I consent >> .env
# Set to True to enable the GW footer
- echo SFM_ENABLE_GW_FOOTER=False >> .env
# To send email, set these correctly.
- echo SFM_SMTP_HOST=smtp.gmail.com >> .env
- echo SFM_EMAIL_USER=someone@gmail.com >> .env
- echo SFM_EMAIL_PASSWORD=password >> .env
# An optional contact email at your institution that is provided to users.
- echo SFM_CONTACT_EMAIL=sfm@yourinstitution.edu >> .env
# To enable connecting to social media accounts, provide the following.
- echo TWITTER_CONSUMER_KEY=mBbq9ruffgEcfsktgQztTHUir8Kn0 >> .env
- echo WEIBO_API_KEY=13132044538 >> .env
- echo TUMBLR_CONSUMER_KEY=Fki09cW957y56h6fhRtCnig14QhpM0pjuHbDWMrZ9aPXcsthVQq >> .env
- echo TUMBLR_CONSUMER_SECRET=aPTpFRE207sV146xB3dif8n8kBb7EpUfWBxuHcB4gfVP >> .env

# For automatically created admin account
- echo SFM_SITE_ADMIN_NAME=sfadmin >> .env
- echo SFM_SITE_ADMIN_EMAIL=nowhere@example.com >> .env
- echo SFM_SITE_ADMIN_PASSWORD=password >> .env

## RABBITMQ CONFIGURATION
- echo RABBITMQ_USER=sfm_user >> .env
- echo RABBITMQ_PASSWORD=password >> .env
- echo RABBITMQ_MANAGEMENT_PORT=15672 >> .env

## DB CONFIGURATION
- echo POSTGRES_PASSWORD=password >> .env

- docker-compose up -d
- docker-compose scale twitterrestharvester=2 twitterpriorityrestharvester=2

When the instance is launched, SFM will be installed and started.

Note the following:
• Starting up the EC2 instance will take several minutes.

• This has been tested with *Ubuntu Server 14.04 LTS*, but may work with other AMI types.

• For suggestions on sizing your SFM server, see *Server sizing*.

• If you need to make additional changes to your `docker-compose.yml`, you can ssh into the EC2 instance and make changes. `docker-compose.yml` and `.env` will be in the default user’s home directory.

• Make sure to configure a security group that exposes the proper ports. To see which ports are used by which services, see example.prod.docker-compose.yml.

• To learn more about configuring EC2 instances with user data, see the AWS user guide.

### 8.4 Configuration

Configuration is documented in `example.env`. For a production deployment, pay particular attention to the following:

• Set new passwords for `SFM_SITE_ADMIN_PASSWORD`, `RABBIT_MQ_PASSWORD`, and `POSTGRES_PASSWORD`.

• The *data volume strategy* is used to manage the volumes that store SFM’s data. By default, normal Docker volumes are used. To use a host volume instead, change the `DATA_VOLUME` and `PROCESSING_VOLUME` settings. Host volumes are recommended for production because they allow access to the data from outside of Docker.

• Set the `SFM_HOSTNAME` and `SFM_PORT` appropriately. These are the public hostname (e.g., sfm.gwu.edu) and port (e.g., 80) for SFM.

• Email is configured by providing `SFM_SMTP_HOST`, `SFM_EMAIL_USER`, and `SFM_EMAIL_PASSWORD`. (If the configured email account is hosted by Google, you will need to configure the account to “Allow less secure apps.” Currently this setting is accessed, while logged in to the google account, via [https://myaccount.google.com/security#connectedapps](https://myaccount.google.com/security#connectedapps).

• Application credentials for social media APIs are configured in by providing the `TWITTER_CONSUMER_KEY`, `TWITTER_CONSUMER_SECRET`, `WEIBO_API_KEY`, `WEIBO_API_SECRET`, and/or `TUMBLR_CONSUMER_KEY`, `TUMBLR_CONSUMER_SECRET`. These are optional, but will make acquiring credentials easier for users. For more information and alternative approaches see *API Credentials*.

• Set an admin email address with `SFM_SITE_ADMIN_EMAIL`. Problems with SFM are sent to this address.

• Set an SFM contact email address with `SFM_CONTACT_EMAIL`. Users are provided with this address.

• For branding in the SFM UI footer, provide `SFM_INSTITUTION_NAME` and `SFM_INSTITUTION_LINK`. (There is also a GW-specific footer available which, when enabled, appears below the standard footer. The GW-specific footer is disabled by default. The environment variable that controls this is `SFM_ENABLE_GW_FOOTER`.)

• To enable the cookie consent popup:
  
  – Set `SFM_ENABLE_COOKIE_CONSENT` to True.
  
  – Optionally, customize the text of `SFM_COOKIE_CONSENT_HTML`. HTML tags are allowed in `SFM_COOKIE_CONSENT_HTML`. For instance, you may wish to use an `<a>` (anchor tag) to include a link to your institution’s privacy policy web page.
  
  – Optionally, customize the wording of the cookie consent button in `SFM_COOKIE_CONSENT_BUTTON_TEXT`. 

---

Chapter 8. Installation and configuration
Note that if you make a change to configuration after SFM is brought up, you will need to restart containers. If the change only applies to a single container, then you can stop the container with `docker kill <container name>`. If the change applies to multiple containers (or you’re not sure), you can stop all containers with `docker-compose stop`. Containers can then be brought back up with `docker-compose up -d` and the configuration change will take effect.

### 8.5 HTTPS

To run SFM with HTTPS:

1. Create or acquire a valid certificate and private key.
2. In `docker-compose.yml` uncomment the nginx-proxy container and set the paths under `volumes` to point to your certificate and key.
3. In `.env` change `USE_HTTPS` to True and `SFM_PORT` to 8080. Make sure that `SFM_HOSTNAME` matches your certificate.
4. Start up SFM.

Note:

- HTTPS will run on 443. Port 80 will redirect to 443.
- For more information on nginx-proxy, including advanced configuration see [https://github.com/jwilder/nginx-proxy](https://github.com/jwilder/nginx-proxy).
- If you receive a 502 (bad gateway), wait until SFM UI has completely started. If the 502 continues, troubleshoot SFM UI.

### 8.6 Stopping

To stop the containers gracefully:

```bash
docker-compose stop -t 180 twitterstreamharvester
docker-compose stop -t 45
```

SFM can then be restarted with `docker-compose up -d`.

### 8.7 Server restarts

If Docker is configured to automatically start when the server starts, then SFM will start. (This is enabled by default when Docker is installed.)

SFM will even be started if it was stopped prior to the server reboot. If you do not want SFM to start, then configure Docker to not automatically start.

To configure whether Docker is automatically starts, see *Stopping Docker from automatically starting*.

### 8.8 Upgrading

Following are general instructions for upgrading SFM versions. Always consult the release notes of the new version to see if any additional steps are required.
1. Stop the containers gracefully:

```bash
docker-compose stop -t 180 twitterstreamharvester
docker-compose stop -t 45
```

This may take several minutes.

2. Make a copy of your existing `docker-compose.yml` and `.env` files:

```bash
cp docker-compose.yml old.docker-compose.yml
cp .env old.env
```

3. Get the latest `example.prod.docker-compose.yml`. If you previously cloned the sfm-docker repository then:

```bash
git pull
# Replace 2.3.0 with the correct version.
git checkout 2.3.0
cp example.prod.docker-compose.yml docker-compose.yml
```

otherwise, replacing 2.3.0 with the correct version:

```bash
curl -L https://raw.githubusercontent.com/gwu-libraries/sfm-docker/2.3.0/example.prod.docker-compose.yml > docker-compose.yml
```

4. If you customized your previous `docker-compose.yml` file, make the same changes in your new `docker-compose.yml`.

5. Make any changes in your `.env` file prescribed by the release notes.

6. Bring up the containers:

```bash
docker-compose up -d
```

It may take several minutes for the images to be downloaded and the containers to start.

7. Deleting images from the previous version is recommended to prevent Docker from filling up too much space. Replacing 2.0.2 with the correct previous version:

```bash
docker rmi $(docker images | grep "2.0.2" | awk '{print $3}')
```

You may also want to periodically clean up Docker (>= 1.13) with `docker system prune`.

### 8.9 Server sizing

While we have not performed any system engineering analysis of optimal server sizing for SFM, the following are different configurations that we use:

<table>
<thead>
<tr>
<th>Use</th>
<th>Server type</th>
<th>Processors</th>
<th>RAM (gb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production</td>
<td></td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Sandbox</td>
<td>m4.large (AWS)</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Use in a class</td>
<td>m4.xlarge (AWS)</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Continuous integration</td>
<td>t2.medium (AWS)</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Heavy dataset processing</td>
<td>m4.4xlarge (AWS)</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td>Development</td>
<td>Docker for Mac</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
There are several mechanisms for monitoring (and troubleshooting) SFM. For more information on troubleshooting, see *Troubleshooting*.

### 9.1 Monitor page

To reach the monitoring page, click “Monitor” on the header of any page in SFM UI.

The monitor page provides status and queue lengths for SFM components, including harvesters and exporters. The status is based on the most recent status reported back by each harvester or exporter (within the last 3 days). A harvester or exporter reports its status when it begins a harvest or export. It also reports its status when it completes the harvest or export. Harvesters will also provide status updates periodically during a harvest.

Note that if there are multiple instances of a harvester or exporter (created with `docker-compose scale`), each instance will be listed.

The queue length lists the number of harvest or export requests that are waiting. A long queue length can indicate that additional harvesters or exporters are needed to handle the load (see *Scaling up with Docker*) or that there is a problem with the harvester or exporter.

The queue length for SFM UI is also listed. This is a queue of status update messages from harvesters or exporters. SFM UI uses these messages to update the records for harvests and exports. Any sort of a queue here indicates a problem.

### 9.2 Logs

It can be helpful to peek at the logs to get more detail on the work being performed by a harvester or exporter.
9.2.1 Docker logs

The logs for harvesters and exporters can be accessed using Docker’s `log` commands.

First, determine the name of the harvester or exporter using `docker ps`. In general, the name will be something like `sfm_twitterrestharvester_1`.

Second, get the log with `docker logs <name>`.

Add `-f` to follow the log. For example, `docker logs -f sfm_twitterrestharvester_1`.

Add `--tail=<number of lines>` to get the tail of the log. For example, `docker logs --tail=100 sfm_twitterrestharvester_1`.

Side note: To follow the logs of all services, use `docker-compose logs -f`.

9.2.2 Twitter Stream Harvester logs

Since the Twitter Stream Harvester runs multiple harvests on the same host, accessing its logs are a bit different.

First, determine the name of the Twitter Stream Harvester and the container id using `docker ps`. The name will probably be `sfm_twitterstreamharvester_1` and the container id will be something like `bffcae5d0603`.

Second, determine the harvest id. This is available from the harvest’s detail page.

Third, get the stdout log with `docker exec -t <name> cat /sfm-data/containers/<container id>/log/<harvest id>.out.log`. To get the stderr log, substitute `.err` for `.out`.

To follow the log, use `tail -f` instead of `cat`. For example, `docker exec -t sfm_twitterstreamharvester_1 tail -f /sfm-data/containers/bffcae5d0603/log/d4493eed5f4f49c6a1981c89cb5d525f.err.log`.

9.3 RabbitMQ management console

The RabbitMQ Admin is usually available on port 15672. For example, `http://localhost:15672/`.
Designated users have access to SFM UI’s Django Admin interface by selecting Welcome > Admin on the top right of the screen. This interface will allow adding, deleting, or changing database records for SFM UI. Some of the most salient uses for this capability are given below.

## 10.1 Managing groups

To allow for multiple users to control a collection set:

1. Create a new group.
2. Add users to the group. (This is done from the user’s admin page, not the group’s admin page.)
3. Assign the collection set to the group. This is done from the collection set detail page or from the collection set admin page.

## 10.2 Deactivating collections

Deactivating a collection indicates that you have completed collecting data for that collection. Deactivated collections will be removed from some of the lists in SFM UI and will not appear in the harvest status emails.

Collections can be deactivated using the “Deactivate” button on the collection detail page.

Note:

- A deactivated collection can be re-activated from the collection detail page. A deactivated collection must be re-activated before it can be edited or turned on.
- A collection set is considered deactivated when it has no active collections. It will also be removed from some of the lists in SFM UI and not appear in harvest status emails.
10.3 Sharing collections

Changing the visibility of a collection to “Other users” will allow the collection to be viewed by all SFM users. The visibility of a collection can be changed by editing the collection.

Note: * A collection set is shared when it has a shared collection. * Shared collection sets will be listed on a separate tab of the collection set list page.

10.4 Deleting items

Records can be deleted using the Admin Interface. It is recommended to minimize deletion; in particular, collections should be turned off and seeds made inactive.

Note the following when deleting:

- Cascades delete, i.e., when a record is deleted any other records that depend on it will also be deleted. Before the deletion is performed, you will be informed what dependent records will be deleted.

- When deleting collection sets, collections, harvests, WARC, and exports the corresponding files will be deleted. Thus, if you delete a collection set all data and metadata will be deleted. **Be careful.**

10.5 Moving collections

Collections can be moved from one collection set to another. This is done by changing the collection set for the collection in the Admin Interface.

Note the following when moving collections:

- The collections files are moved as well, as the directory structure includes the collection set’s identifier.

- The path for WARC files in WARC records are updated.

- Make sure harvesting is turned off and all harvests and exports are completed before moving.

- Previous exports will become unavailable after the move.

10.6 Allowing access to Admin Interface

To allow a user to have access to the Admin Interface, give the user Staff status or Superuser status. This is done from the user’s admin page.
11.1 Accounts

Social Feed Manager allows users to self-sign up for accounts. Those accounts are stored and managed by SFM. Future versions of SFM may support authentication against external systems, e.g., Shibboleth.

Because creating accounts is not restricted, it is encouraged to take other measures to secure SFM such as restricting access to the IP range of your institution.

11.2 Groups

By default, a group is created for each user and the user is placed in group. To create additional groups and modify group membership use the Admin interface.

In general, users and groups can be administered from the Admin interface.

11.3 Authorizations

A collection set is owned by a group. The members of that group can administer the collection set and all of its collections. Thus, to enable a team to collaborate on a collection set, create a group that includes all of the team members and assign ownership of the team to the group.

When the visibility on a collection is set to “Other users” then any user that is logged into SFM can view the collection and request an export. (The user cannot make a change to the collection.)

A user that is designated as staff can view all collections.

A user that is designated as a superuser can administer all collections.
This page contains information about Docker that is useful for installation, administration, and development.

12.1 Installing Docker

Docker Engine and Docker Compose

On OS X:

- Install Docker for Mac.
- If you are using Docker Toolbox, switch to Docker for Mac.

On Ubuntu:

- If you have difficulties with the `apt` install, try the `pip` install.
- The docker group is automatically created. Adding your user to the docker group avoids having to use sudo to run docker commands. Note that depending on how users/groups are set up, you may need to manually need to add your user to the group in `/etc/group`.

While Docker is available on other platforms (e.g., Windows, Red Hat Enterprise Linux), the SFM team does not have any experience running SFM on those platforms.

12.2 Helpful commands

`docker-compose up -d` Bring up all of the containers specified in the docker-compose.yml file. If a container has not yet been pulled, it will be pulled. If a container has not yet been built it will be built. If a container has been stopped (“killed”) it will be re-started. Otherwise, a new container will be created and started (“run”).

`docker-compose pull` Pull the latest images for all of the containers specified in the docker-compose.yml file with the `image` field.
**docker-compose build** Build images for all of the containers specified in the docker-compose.yml file with the `build` field. Add `--no-cache` to re-build the entire image (which you might want to do if the image isn’t building as expected).

**docker ps** List running containers. Add `–a` to also list stopped containers.

**docker-compose kill** Stop all containers.

**docker kill <container name>** Stop a single container.

**docker-compose rm -v --force** Delete the containers and volumes.

**docker rm -v <container name>** Delete a single container and volume.

**docker rm $(docker ps -a -q) -v** Delete all containers.

**docker-compose logs** List the logs from all containers. Add `–f` to follow the logs.

**docker logs <container name>** List the log from a single container. Add `–f` to follow the logs.

**docker-compose -f <docker-compose.yml filename> <command>** Use a different docker-compose.yml file instead of the default.

**docker exec -it <container name> /bin/bash** Shell into a container.

**docker rmi <image name>** Delete an image.

**docker rmi $(docker images -q)** Delete all images

**docker-compose scale <service name>=<number of instances>** Create multiple instances of a service.

### 12.3 Scaling up with Docker

Most harvesters and exporters handle one request at a time; requests for exports and harvests queue up waiting to be handled. If requests are taking too long to be processed you can scale up (i.e., create additional instances of) the appropriate harvester or exporter.

To create multiple instances of a service, use **docker-compose scale**.

The harvester most likely to need scaling is the Twitter REST harvester since some harvests (e.g., broad Twitter searches) may take a long time. To scale up the Twitter REST harvester to 3 instances use:

```
docker-compose scale twitterrestharvester=3
```

To spread containers across multiple containers, use **Docker Swarm**.

Using compose in production provides some additional guidance.

### 12.4 Stopping Docker from automatically starting

Docker automatically starts when the server starts. To control this:

#### 12.4.1 Ubuntu 14 (Upstart)

Stop Docker from automatically starting:

```
echo manual | sudo tee /etc/init/docker.override
```
Allow Docker to automatically start:

```
sudo rm /etc/init/docker.override
```

Manually start Docker:

```
sudo service docker start
```

### 12.4.2 Ubuntu 16 (Systemd)

Stop Docker from automatically starting:

```
sudo systemctl disable docker
```

Allow Docker to automatically start:

```
sudo systemctl enable docker
```

Manually start Docker:

```
sudo systemctl start docker
```
13.1 Overview

Collections and collection sets are portable. That means they can be moved to another SFM instance or to another environment, such as a repository. This can also be used to backup an SFM instance.

A collection includes all of the social media items (stored in WARC) and the database records for the collection sets, collections, users, groups, credentials, seeds, harvests, and WARC, as well as the history of collection sets, collections, credentials, and seeds. The database records are stored in JSON format in the records subdirectory of the collection. Each collection has a complete set of JSON database records to support loading it into a different SFM instance.

Here are the JSON database records for an example collection:

```
[root@1da93af43b5:/sfm-data/collection_set/4c59ebf2dcdc4a0e9660e32d004fa846/]
  collection.json  groups.json  historical_collection.json  historical_seeds.
  ->  users.json
  collection_set.json  harvest_stats.json  historical_collection_set.json  info.json
  ->  warcs.json
  credentials.json  harvests.json  historical_credentials.json  seeds.json
```

Thus, moving a collection set only requires moving/copying the collection set’s directory; moving a collection only requires moving/copying a collection’s directory. Collection sets are in /sfm-data/collection_set and are named by their collection set ids. Collections are subdirectories of their collection set and are named by their collection ids.

A README.txt is automatically created for each collection and collection set. Here a README.txt for an example collection set:

```
This is a collection set created with Social Feed Manager.

Collection set name: test collection set
Collection set id: 4c59ebf2dcdc4a0e9660e32d004fa846
```

(continues on next page)
This collection set contains the following collections:
* test twitter sample (collection id 59f9ff647ff4fa28fd7e5bc4d161743)
* test twitter user timeline (collection id 072ff07ea9954b39a1883e979de92d22)

Each of these collections contains a README.txt.

Updated on Oct. 18, 2016, 3:09 p.m.

13.2 Preparing to move a collection set / collection

Nothing needs to be done to prepare a collection set or collection for moving. The collection set and collection directories contain all of the files required to load it into a different SFM instance.

The JSON database records are refreshed from the database on a nightly basis. Alternatively, they can be refreshed used the serializecollectionset and serializecollection management commands:

```bash
root@1da93afd43b5:/opt/sfm-ui/sfm# ./manage.py serializecollectionset 4c59ebf2d
```

13.3 Loading a collection set / collection

1. Move/copy the collection set/collection to /sfm-data/collection_set. Collection sets should be placed in this directory. Collections should be placed into a collection set directory.

2. Execute the deserializecollectionset or deserializecollection management command:

```bash
root@1da93afd43b5:/opt/sfm-ui/sfm# ./manage.py deserializecollectionset /sfm-data/collection_set/4c59ebf2dc4a0e9660e32d004fa886
```

Note:

- If loading a collection set, all of the collection set’s collections will also be loaded.
- When loading, all related items are also loaded. For example, when a collection is loaded, all of the seeds, harvests, credentials, and their histories are also loaded.
- If a database record already exists for a collection set, loading will not continue for the collection set or any of its collections or related records (e.g., groups).
- If a database record already exists for a collection, loading will not continue for the collection or any of the related records (e.g., users, harvests, WARC).
- If a database record already exists for a user or group, it will not be loaded.
- Collections that are loaded are turned off.
- Users that are loaded are set to inactive.
- A history note is added to collection sets and collections to document the load.
13.4 Moving an entire SFM instance

1. Stop the source instance: `docker-compose stop`.
2. Copy the `/sfm-data` directory from the source server to the destination server.
3. If preserving processing data, also copy the `/sfm-processing` directory from the source server to the destination server.
4. Copy the `docker-compose.yml` and `.env` files from the source server to the destination server.
5. Make any changes necessary in the `.env` file, e.g., `SFM_HOSTNAME`.
6. Start the destination instance: `docker-compose up -d`.

If moving between AWS EC2 instances and `/sfm-data` is on a separate EBS volume, the volume can be detached from the source EC2 instances and attached to the destination EC2 instance.
14.1 Storage volumes

SFM stores data on 2 volumes:

- sfm-data: The data volume is where SFM stores the harvested social media content, the db files, and exports. This is described in more detail below. It is available within containers as /sfm-data.

- sfm-processing: The processing volume is where processed data is stored when using a processing container. (See Command-line exporting/processing.) It is available within containers as /sfm-processing.

14.2 Volume types

There are 2 types of volumes:

- Internal to Docker. The files on the volume will only be available from within Docker containers.

- Linked to a host location. The files on the volumes will be available from within Docker containers and from the host operating system.

The type of volume is specified in the .env file. When selecting a link to a host location, the path on the host environment must be specified:

```bash
# Docker internal volume
DATA_VOLUME=/sfm-data
# Linked to host location
#DATA_VOLUME=/src/sfm-data:/sfm-data
# Docker internal volume
PROCESSING_VOLUME=/sfm-processing
# Linked to host location
#PROCESSING_VOLUME=/src/sfm-processing:/sfm-processing
```
We recommend that you use an internal volume only for development; for other uses linking to a host location is recommended. This makes it easier to place the data on specific storage devices (e.g., NFS or EBS) and to backup the data.

14.3 File ownership

SFM files are owned by the sfm user (default uid 990) in the sfm group (default gid 990). If you use a link to a host location and list the files, the uid and gid may be listed instead of the user and group names.

If you shell into a Docker container, you will be the root user. Make sure that any operations you perform will not leave behind files that do not have appropriate permissions for the sfm user.

Note then when using Docker for Mac and linking to a host location, the file ownership may not appear as expected.

14.4 Directory structure of sfm-data

The following is an outline of the structure of sfm-data:

```
/sfm-data/
  collection_set/
    <collection set id>/
      README.txt (README for collection set)
    <collection id>/
      README.txt (README for collection)
      state.json (Harvest state record)
      records/
        JSON records for the collection metadata
      <year>/<month>/<day>/<hour>/
        WARC files
    containers/
      <container id>/
        Working files for individual containers
    export/
      <export id>/
        Export files
    postgresql/
      Postgres db files
```

14.5 Space warnings

SFM will monitor free space on sfm-data and sfm-processing. Administrators will be notified when the amount of free space crosses a configurable threshold. The threshold is set in the .env file:

```
# sfm-data free space threshold to send notification emails, only ends with MB, GB, TB.
# eg. 500MB, 10GB, 1TB
DATA_VOLUME_THRESHOLD=10GB

# sfm-processing free space threshold to send notification emails, only ends with MB, GB, TB.
# eg. 500MB, 10GB, 1TB
PROCESSING_VOLUME_THRESHOLD=10GB
```
14.6 Moving from a Docker internal volume to a linked volume

These instructions are for Ubuntu. They may need to be adjusted for other operating systems.

1. Stop docker containers:
   
   `docker-compose stop`

2. Copy sfm-data contents from inside the container to a linked volume:
   
   `sudo docker cp sfm_data_1:/sfm-data /

3. Set ownership:
   
   `sudo chown -R 990:990 /sfm-data/*`

4. Change .env:
   
   `#DATA_VOLUME=/sfm-data`
   
   `DATA_VOLUME=/sfm-data:/sfm-data`

5. Restart containers:
   
   `docker-compose up -d`
To make sure you have the best possible experience with SFM, you should be aware of the limitations and known issues:

- Changes to the hostname of server (e.g., from the reboot of an AWS EC2 instance) are not handled (Ticket 435). See also Troubleshooting

- The README file when downloaded and opened in Notepad (Early versions of Windows 10 or below) lacks linebreaks since Notepad cannot read the linebreaks specified within the code. This issue was fixed by Microsoft in the Windows 10 version 1809 (October 2018) of Notepad and the file opens up in the expected format (Ticket 1002).

  If you are using early versions of Windows 10 or below use Windows WordPad to open the README file since it renders the file in the correct format with appropriate linebreaks.

For a complete list of tickets, see [https://github.com/gwu-libraries/sfm-ui/issues](https://github.com/gwu-libraries/sfm-ui/issues)

In addition, you should be aware of the following:

- Access to the Weibo API is limited, so make sure you understand what can be collected.

- SFM does not currently provide a web interface for “replaying” the collected social media or web content.
16.1 General tips

- Upgrade to the latest version of Docker and Docker-Compose.
- Make sure expected containers are running with `docker ps`.
- Check the logs with `docker-compose logs` and `docker logs <container name>`.
- Additional information is available via the admin interface that is not available from the UI. To access the admin interface, log in as an account that has superuser status and under “Welcome, <your name>,” click Admin. By default, a superuser account called `sfmadmin` is created. The password can be found in `.env`.

16.2 Specific problems

16.2.1 Skipped harvests

A new harvest will not be requested if the previous harvest has not completed. Instead, a harvest record will be created with the status of skipped. Some of the reasons that this might happen include:

- Harvests are scheduled too closely together, such that the previous harvest cannot complete before the new harvest is requested.
- There are not enough running harvesters, such that harvest requests have to wait too long before being processed.
- There is a problem with harvesters, such that they are not processing harvest requests.
- Something else has gone wrong, and a harvest request was not completed.

After correcting the problem to resume harvesting for a collection, void the last (non-skipped) harvest. To void a harvest, go to that harvest’s detail page and click the void button.
16.2.2 Connection errors when harvesting

If harvests from a container fail with something like:

```
HTTPSConnectionPool(host='api.flickr.com', port=443): Max retries exceeded with url: /...
...getInfo&format=json (Caused by ProxyError('Cannot connect to proxy.', error('Tunnel...
...connection failed: 500 [Errno -3] Temporary failure in name resolution',))))
```

then stop and restart the container. For example:

```
docker-compose stop flickrharvester
docker-compose up -d
```

16.2.3 Bind error

If when bringing up the containers you receive something like:

```
ERROR: driver failed programming external connectivity on endpoint docker_sfmuiapp_1...
...failed: Port 0.0.0:8080 failed: Port is already allocated
```

it means another application is already using a port configured for SFM. Either shut down the other application or choose a different port for SFM. (Chances are the other application is Apache.)

16.2.4 Bad Request (400)

If you receive a Bad Request (400) when trying to access SFM, your `SFM_HOSTNAME` environment variable is not configured correctly. Check what `SFM_HOSTNAME` is set to in `.env`, and update and restart (docker-compose stop ui then docker-compose up -d) if necessary. For more information, see ALLOWED_HOSTS.

16.2.5 Social Network Login Failure for Twitter

If you receive a Social Network Login Failure when trying to connect a Twitter account, make sure that the Twitter app from which you got the Twitter credentials is configured with a callback URL. The URL should be `http://<SFM hostname>/accounts/twitter/login/callback/`.

If you have made a change to the credentials configured in `.env`, try deleting twitter from Social Applications in the admin interface and restarting SFM UI (docker-compose stop ui then docker-compose up -d).

16.2.6 Docker problems

If you are having problems bringing up the Docker containers (e.g., `driver failed programming external connectivity on endpoint`), restart the Docker service. On Ubuntu, this can be done with:

```
service docker stop
docker stop/waiting

service docker status
docker stop/waiting

service docker start
docker start/running, process 15039
```
16.2.7 CSV export problems

Excel for Mac has problems with unicode characters in CSV files. As a work-around, export to Excel (XLSX) format.

16.3 Still stuck?

Contact the SFM team. We’re happy to help.
17.1 Setting up a development environment

SFM is composed of a number of components. Development can be performed on each of the components separately. For SFM development, it is recommended to run components within a Docker environment (instead of directly in your OS, without Docker).

17.1.1 Step 1: Install Docker and Docker Compose

See *Installing Docker*.

17.1.2 Step 2: Clone sfm-docker and create copies of docker-compose files

For example:

```bash
git clone https://github.com/gwu-libraries/sfm-docker.git
cd sfm-docker
cp example.docker-compose.yml docker-compose.yml
cp example.env .env
```

For the purposes of development, you can make changes to `docker-compose.yml` and `.env`. This will be described more below.

17.1.3 Step 3: Clone the component repos

For example:

```bash
git clone https://github.com/gwu-libraries/sfm-ui.git
```
Repeat for each of the components that you will be working on. Each of these should be in a sibling directory of sfm-docker.

### 17.2 Running SFM for development

To bring up an instance of SFM for development, change to the sfm-docker directory and execute:

```bash
docker-compose up -d
```

You may not want to run all of the containers. To omit a container, simply comment it out in `docker-compose.yml`.

By default, the code that has been committed to master for each of the containers will be executed. To execute your local code (i.e., the code you are editing), you will want to link in your local code. To link in the local code for a container, uncomment the volume definition that points to your local code. For example:

```yaml
volumes:
  - "../sfm-twitter-harvester:/opt/sfm-twitter-harvester"
```

sfm-utils and warcprox are dependencies of many components. By default, the code that has been committed to master for sfm-utils or warcprox will be used for a component. To use your local code as a dependency, you will want to link in your local code. Assuming that you have cloned sfm-utils and warcprox, to link in the local code as a dependency for a container, change `SFM_REQS` in `.env` to “dev” and comment the volume definition that points to your local code. For example:

```yaml
volumes:
  - "../sfm-twitter-harvester:/opt/sfm-twitter-harvester"
  - "../sfm-utils:/opt/sfm-utils"
  - "../warcprox:/opt/warcprox"
```

Note: * As a Django application, SFM UI will automatically detect code changes and reload. Other components must be killed and brought back up to reflect code changes.

### 17.3 Running tests

#### 17.3.1 Unit tests

Some components require a `test_config.py` file that contains credentials. For example, sfm-twitter-harvester requires a `test_config.py` containing:

```python
TWITTER_CONSUMER_KEY = "EHdoTk5Bfg1P5nUa1Efhao"
TWITTER_CONSUMER_SECRET = "StUpemtBxf2cEmaqly52Dd343ihFu9PAiLebuM0mqNOQtXeAlen"
TWITTER_ACCESS_TOKEN = "411876914-c2y2jbk1np025MWEFYnYQK5QNFGBXd8T4k90YkJ1"
TWITTER_ACCESS_TOKEN_SECRET = "jK9QOm5VRF5mfgAN6KgfmCKReqThXVQ1G6qQg8BCeJvp"
```

Note that if this file is not present, unit tests that require it will be skipped. Each component’s README will describe the `test_config.py` requirements.

Also note that some unit tests may fail unless the local environment contains an `LC_ALL` environment variable set to `en_US.UTF-8`.

Unit tests for most components can be run with:
The notable exception is SFM UI, which can be run with:

```bash
cd sfm
./manage.py test --settings=sfm.settings.test_settings
```

### 17.3.2 Integration tests

Many components have integration tests, which are run inside docker containers. These components have a `ci.docker-compose.yml` file which can be used to bring up a minimal environment for running the tests.

As described above, some components require a `test_config.py` file.

To run integration tests, bring up SFM:

```bash
docker-compose -f docker/dev.docker-compose.yml up -d
```

Run the tests:

```bash
docker exec docker_sfmtwitterstreamharvester_1 python -m unittest discover
```

You will need to substitute the correct name of the container. (`docker ps` will list the containers.)

And then clean up:

```bash
docker-compose -f docker/dev.docker-compose.yml kill
docker-compose -f docker/dev.docker-compose.yml rm -v --force
```

For reference, see each component’s `.travis.yml` file which shows the steps of running the integration tests.

### 17.3.3 Smoke tests

sfm-docker contains some smoke tests which will verify that a development instance of SFM is running correctly.

To run the smoke tests, first bring up SFM:

```bash
docker-compose -f example.docker-compose.yml -f smoketests.docker-compose.yml up -d
```

wait, and then run the tests:

```bash
docker-compose -f example.docker-compose.yml -f smoketests.docker-compose.yml run --rm smoketests /bin/bash -c "appdeps.py --port-wait mq:5672 --port-wait ui:8080 && python -m unittest discover"
```

Note that the smoke tests are not yet complete and require test fixtures that are only available in a development deploy.

For reference, the continuous integration deploy instructions shows the steps of running the smoke tests.

### 17.4 Requirements files

This will vary depending on whether a project has warcprox and sfm-utils as a dependency, but in general:

- `requirements/common.txt` contains dependencies, except warcprox and sfm-utils.
• requirements/release.txt references the last released version of warcprox and sfm-utils.
• requirements/master.txt references the master version of warcprox and sfm-utils.
• requirements/dev.txt references local versions of warcprox and sfm-utils in development mode.

To get a complete set of dependencies, you will need common.txt and either release.txt, master.txt or dev.txt. For example:

```
virtualenv ENV
source ENV/bin/activate
pip install -r requirements/common.txt -r requirements/dev.txt
```

## 17.5 Development tips

### 17.5.1 Admin user accounts

Each component should automatically create any necessary admin accounts (e.g., a django admin for SFM UI). Check .env for the username/passwords for those accounts.

### 17.5.2 RabbitMQ management console

The RabbitMQ management console can be used to monitor the exchange of messages. In particular, use it to monitor the messages that a component sends, create a new queue, bind that queue to `sfm_exchange` using an appropriate routing key, and then retrieve messages from the queue.

The RabbitMQ management console can also be used to send messages to the exchange so that they can be consumed by a component. (The exchange used by SFM is named `sfm_exchange`.)

For more information on the RabbitMQ management console, see `RabbitMQ`.

### 17.5.3 Blocked ports

When running on a remote VM, some ports (e.g., 15672 used by the RabbitMQ management console) may be blocked. SSH port forwarding can help make those ports available.

### 17.5.4 Django logs

Django logs for SFM UI are written to the Apache logs. In the docker environment, the level of various loggers can be set from environment variables. For example, setting `SFM_APSCHEDULER_LOG` to `DEBUG` in the `docker-compose.yml` will turn on debug logging for the apscheduler logger. The logger for the SFM UI application is called `ui` and is controlled by the `SFM_UI_LOG` environment variable.

### 17.5.5 Apache logs

In the SFM UI container, Apache logs are sent to stdout/stderr which means they can be viewed with `docker-compose logs` or `docker logs <container name or id>`.
17.5.6 Initial data

The development and master docker images for SFM UI contain some initial data. This includes a user (“testuser”, with password “password”). For the latest initial data, see fixtures.json. For more information on fixtures, see the Django docs.

17.5.7 Runserver

There are two flavors of the development docker image for SFM UI. gwul/sfm-ui:master runs SFM UI with Apache, just as it will in production. gwul/sfm-ui:master-runserver runs SFM UI with runserver, which dynamically reloads changed Python code. To switch between them, change UI_TAG in .env.

Note that as an byproduct of how runserver dynamically reloads Python code, there are actually 2 instances of the application running. This may produce some odd results, like 2 schedulers running. This will not occur with Apache.

17.5.8 Job schedule intervals

To assist with testing and development, a 5 minute interval can be added by setting SFM_FIVE_MINUTE_SCHEDULE to True in the docker-compose.yml.

17.5.9 Connecting to the database

To connect to postgres using psql:

```
docker exec -it sfm_db_1 psql -h db -U postgres -d sfmdatabase
```

You will be prompted for the password, which you can find in .env.

17.6 Docker tips

17.6.1 Building vs. pulling

Containers are created from images. Images are either built locally or pre-built and pulled from Docker Hub. In both cases, images are created based on the docker build (i.e., the Dockerfile and other files in the same directory as the Dockerfile).

In a docker-compose.yml, pulled images will be identified by the image field, e.g., image: gwul/sfm-ui:master. Built images will be identified by the build field, e.g., build: app-dev.

In general, you will want to use pulled images. These are automatically built when changes are made to the Github repos. You should periodically execute docker-compose pull to make sure you have the latest images.

You may want to build your own image if your development requires a change to the docker build (e.g., you modify fixtures.json).

17.6.2 Killing, removing, and building in development

Killing a container will cause the process in the container to be stopped. Running the container again will cause process to be re-started. Generally, you will kill and run a development container to get the process to be run with changes you’ve made to the code.
Removing a container will delete all of the container’s data. During development, you will remove a container to make sure you are working with a clean container.

Building a container creates a new image based on the Dockerfile. For a development image, you only need to build when making changes to the docker build.
CHAPTER 18

Writing a harvester

18.1 Requirements

• Implement the *Messaging Specification* for harvesting social media content. This describes the messages that must be consumed and produced by a harvester.

• Write harvested social media to a WARC, following all relevant guidelines and best practices. The message for announcing the creation of a WARC is described in the Messaging Specification. The WARC file must be written to `<base path>/<harvest year>/<harvest month>/<harvest day>/<harvest hour>/`, e.g., `/data/test_collection_set/2015/09/12/19/`. (Base path is provided in the harvest start message.) Any filename may be used but it must end in `.warc` or `.warc.gz`. It is recommended that the filename include the harvest id (with file system unfriendly characters removed) and a timestamp of the harvest.

• Extract urls for related content from the harvested social media content, e.g., a photo included in a tweet. The message for publishing the list of urls is described in the Messaging Specification.

• Document the harvest types supported by the harvester. This should include the identifier of the type, the API methods called, the required parameters, the optional parameters, what is included in the summary, and what urls are extracted. See the Flickr Harvester as an example.

• The smoke tests must be able to prove that a harvester is up and running. At the very least, the smoke tests should check that the queues required by a harvester have been created. (See `test_queues()`.)

• Be responsible for its own state, e.g., keeping track of the last tweet harvested from a user timeline. See `sfmutils.state_store` for re-usable approaches to storing state.

• Create all necessary exchanges, queues, and bindings for producing and consuming messages as described in *Messaging*.

• Provide master and production Docker images for the harvester on Docker Hub. The master image should have the `master` tag and contain the latest code from the master branch. (Setup an automated build to simplify updating the master image.) There must be a version specific production images, e.g., `1.3.0` for each release. For example, see the Flickr Harvester's dockerfiles and Docker Hub repo.
18.2 Suggestions

- See sfm-utils for re-usable harvester code. In particular, consider subclassing BaseHarvester.
- Create a development Docker image. The development Docker images links in the code outside of the container so that a developer can make changes to the running code. For example, see the Flickr harvester development image.
- Create a development docker-compose.yml. This should include the development Docker image and only the additional images that the harvester depends on, e.g., a Rabbit container. For example, see the Flickr harvester development docker-compose.yml.
- When possible, use existing API libraries.
- Consider write integration tests that test the harvester in an integration test environment. (That is, an environment that includes the other services that the harvester depends on.) For example, see the Flickr Harvester’s integration tests.
- See the Twitter harvester unit tests for a pattern on configuring API keys in unit and integration tests.

18.3 Notes

- Harvesters can be written in any programming language.
- Changes to gwu-libraries/* repos require pull requests. Pull requests are welcome from non-GWU developers.
CHAPTER 19

Messaging

19.1 RabbitMQ

RabbitMQ is used as a message broker.

The RabbitMQ management console is exposed at http://<your docker host>:15672/. The username is sfm_user. The password is the value of RABBITMQ_DEFAULT_PASS in secrets.env.

19.2 Publishers/consumers

- The hostname for RabbitMQ is mq and the port is 5672.
- It cannot be guaranteed that the RabbitMQ docker container will be up and ready when any other container is started. Before starting, wait for a connection to be available on port 5672 on rabbit. See appdeps.py for docker application dependency support.
- Publishers/consumers may not assume that the requisite exchanges/queues/bindings have previously been created. They must declare them as specified below.

19.3 Exchange

sfm_exchange is a durable topic exchange to be used for all messages. All publishers/consumers must declare it:

```
#Declare sfm_exchange
from kombu import Connection

exchange = Exchange(name="sfm_exchange,
          type="topic", durable=True)
exchange(channel).declare()
```
19.4 Queues

All queues must be declared durable:

```python
from kombu import Queue

queue = Queue(name="harvester",
              exchange=exchange,
              channel=channel,
              durable=True)

queue.declare()
queue.bind_to(exchange=exchange,
              routing_key="harvest.status.*.*")
```
20.1 Introduction

SFM is architected as a number of components that exchange messages via a messaging queue. To implement functionality, these components send and receive messages and perform certain actions. The purpose of this document is to describe this interaction between the components (called a “flow”) and to specify the messages that they will exchange.

Note that as additional functionality is added to SFM, additional flows and messages will be added to this document.

20.2 General

- Messages may include extra information beyond what is specified below. Message consumers should ignore any extra information.
- RabbitMQ will be used for the messaging queue. See the Messaging docs for additional information. It is assumed in the flows below that components receive messages by connecting to appropriately defined queues and publish messages by submitting them to the appropriate exchange.

20.3 Harvesting social media content

Harvesting is the process of retrieving social media content from the APIs of social media services and writing to WARC files.

20.3.1 Background information

- A requester is an application that requests that a harvest be performed. A requester may also want to monitor the status of a harvest. In the current architecture, the SFM UI serves the role of requester.
• A stream harvest is a harvest that is intended to continue indefinitely until terminated. A harvest of a Twitter sample stream is an example of a stream harvest. A stream harvest is different from a non-stream harvest in that a requester must both start and optionally stop a stream harvest. Following the naming conventions from Twitter, a harvest of a REST, non-streaming API will be referred to as a REST harvest.

• Depending on the implementation, a harvester may produce a single warc or multiple warcs. It is likely that in general stream harvests will result in multiple warcs, but REST harvest will result in a single warc.

20.3.2 Flow

The following is the flow for a harvester performing a REST harvest and creating a single warc:

1. Requester publishes a harvest start message.
2. Upon receiving the harvest message, a harvester:
   1. Makes the appropriate api calls.
   2. Writes the api calls to a warc.
3. Upon completing the api harvest, the harvester:
   1. Publishes a warc created message.
   2. Publishes a harvest status message with the status of *completed success* or *completed failure*.

The following is the message flow for a harvester performing a stream harvest and creating multiple warcs:

1. Requester publishes a harvest start message.
2. Upon receiving the harvest message, a harvester:
   1. Opens the api stream.
   2. Writes the stream results to a warc.
3. When rotating to a new warc, the harvester publishes a warc created message.
4. At intervals during the harvest, the harvester:
   1. Publishes a harvest status message with the status of *running*.
5. When ready to stop, the requester publishes a harvest stop message.
6. Upon receiving the harvest stop message, the harvester:
   1. Closes the api stream.
   2. Publishes a final warc created message.
   3. Publishes a final harvest status message with the status of *completed success* or *completed failure*.

• Any harvester may send harvest status messages with the status of *running* before the final harvest status message. A harvester performing a stream harvest must send harvest status messages at regular intervals.

• A requester should not send harvest stop messages for a REST harvest. A harvester performing a REST harvest may ignore harvest stop messages.

20.3.3 Messages

**Harvest start message**

Harvest start messages specify for a harvester the details of a harvest. Example:
Another example:

```json
{
  "id": "test:1",
  "type": "twitter_search",
  "path": "/sfm-data/collections/3989a5f99e41487aaef698680537c3f5/6980fac666c54322a2ebdcb2a9510f5",
  "seeds": [
    {
      "id": "32786222ef374eb38f1c5d56321c99e8",
      "token": "gwu",
      "uid": "0e789cdd0f6b41b5950f569676702182",
      "token": "gelman"
    }
  ],
  "credentials": {
    "consumer_key": "EHde7ksBGgf1bP5nUa1Efhaeo",
    "consumer_secret": "ZtUpemtBkf2maqF1y52D5dihFPAlEbuM0mqN0jeQtXeAlen",
    "access_token": "48186914-c2yZjgbk13npUZ5MWEFQKQSNFBXd8T9r4k90YkJJ",
    "access_token_secret": "jK9QOmn5Vbbmfg2ANT6KgfmRRqV8ThXVQL6qQ8Bejvp"
  },
  "collection_set": {
    "id": "3989a5f99e41487aaef698680537c3f5"
  }
}
```
• The routing key will be harvest.start.<social media platform>.<type>. For example, harvest.start.flickr.flickr_photo.

• **id**: A globally unique identifier for the harvest, assigned by the requester.

• **type**: Identifies the type of harvest, including the social media platform. The harvester can use this to map to the appropriate api calls.

• **seeds**: A list of seeds to harvest. Each seed is represented by a map containing **id**, **token** and (optionally) **uid**. Note that some harvest types may not have seeds.

• **options**: A name/value map containing additional options for the harvest. The contents of the map are specific to the type of harvest. (That is, the seeds for a flickr photo are going to be different than the seeds for a twitter user timeline.)

• **credentials**: All credentials that are necessary to access the social media platform. Credentials is a name/value map; the contents are specific to a social media platform.

• **path**: The base path for the collection.

### Harvest stop message

Harvest stop messages tell a harvester perform a stream harvest to stop. Example:

```json
{
  "id": "sfmui:45"
}
```

• The routing key will be harvest.stop.<social media platform>.<type>. For example, harvest.stop.twitter.filter.

### Harvest status message

Harvest status messages allow a harvester to provide information on the harvests it performs. Example:

```json
{
  "id": "sfmui:45",
  "status": "completed success",
  "date_started": "2015-07-28T11:17:36.640044",
  "date_ended": "2015-07-28T11:17:42.539470",
  "infos": [],
  "warnings": [],
  "errors": [],
  "stats": {
    "2016-05-20": {
      "photos": 12,
    },
    "2016-05-21": {
      "photos": 19,
    },
  },
  "token_updates": {
```

(continues on next page)
The routing key will be harvest.status.<social media platform>.<type>. For example, harvest.status.flickr.flickr_photo.

status: Valid values are completed success, completed failure, or running.

infos, warnings, and errors: Lists of messages. A message should be an object (i.e., dictionary) containing a code and message entry. It may optionally contain a seed_id entry giving the seed id to which the messages applies. Codes should be consistent to allow message consumers to identify types of messages.

stats: A count of items that are harvested by date. Items should be a human-understandable labels (plural and lower-cased). Stats is optional for in progress statuses, but required for final statuses.

token_updates: A map of uids to tokens for which a token change was detected while harvesting. For example, for Twitter a token update would be provided whenever a user’s screen name changes.

uids: A map of tokens to uids for which a uid was identified while harvesting at not provided in the harvest start message. For example, for Flickr a uid would be provided containing the NSID for a username.

warcs.'count': The total number of WARC’s created during this harvest.

warcs.'bytes': The total number of bytes of the WARC’s created during this harvest.

service, host, and instance identify what performed the harvest. service is the name of the harvester. host is the Docker container id. instance is the harvest process identifier (PID) within the container. This is useful in cases where there are multiple instances of a service on a host.

Warc created message

Warc created message allow a harvester to provide information on the warcs that are created during a harvest. Example:

```json
{
    "warc": {
        "sha1": "7512ee1c227c29332172118f0b79b2ca75cbe8979",
        "bytes": 26146,
        "id": "aba6033aafce4fbbad846026ca47f13e",
        "date_created": "2015-07-28T11:17:36.640178"
    },
    "collection_set": {
        "id": "3989a5f99e41487aaef698680537c3f5"
    },
    "collection": {
```
• The routing key will be `warc_created`.
• Each warc created message will be for a single warc.

## 20.4 Exporting social media content

Exporting is the process of extracting social media content from WARCs and writing to export files. The exported content may be a subset or derivate of the original content. A number of different export formats will be supported.

### 20.4.1 Background information

- A requester is an application that requests that an export be performed. A requester may also want to monitor the status of an export. In the current architecture, the SFM UI serves the role of requester.
- Depending on the nature of the export, a single or multiple files may be produced.

### 20.4.2 Flow

The following is the flow for an export:

1. Requester publishes an export start message.
2. Upon receiving the export start message, an exporter:
   1. Makes calls to the SFM REST API to determine the WARC files from which to export.
   2. Limits the content is specified by the export start message.
   3. Writes to export files.
3. Upon completing the export, the exporter publishes an export status message with the status of `completed success` or `completed failure`.

### Export start message

Export start messages specify the requests for an export. Example:

```json
{
    "id": "f3ddcbfc5d6b43139d04d680d278852e",
    "type": "flickr_user",
    "collection": {
        "id": "005b131f5f854402afa2b08a4b7ba960"
    },
    "path": "/sfm-data/exports/45",
    "format": "csv"
}
```
Another example:

```json
{
  "id": "f3ddcbfc5d6b43139d04d680d278852e",
  "type": "flickr_user",
  "seeds": [
    {
      "id": "48722ac6154241f592fd74da775b7ab7",
      "uid": "23972344@N05"
    },
    {
      "id": "3ce76759a3ee40b894562a35359dfa54",
      "uid": "85779209@N08"
    }
  ],
  "path": "/sfm-data/exports/45",
  "format": "json",
  "segment_size": null
}
```

- The routing key will be `export.start.<social media platform>.<type>`. For example, `export.start.flickr.flickr_user`.
- `id`: A globally unique identifier for the harvest, assigned by the requester.
- `type`: Identifies the type of export, including the social media platform. The export can use this to map to the appropriate export procedure.
- `seeds`: A list of seeds to export. Each seed is represented by a map containing `id` and `uid`.
- `collection`: A map containing the `id` of the collection to export.

Each export start message must have a `seeds` or `collection` but not both.
- `path`: A directory into which the export files should be placed. The directory may not exist.
- `format`: A code for the format of the export. (Available formats may change.)
- `dedupe`: If true, duplicate social media content should be removed.
- `item_date_start` and `item_date_end`: The date of social media content should be within this range.
- `harvest_date_start` and `harvest_date_end`: The harvest date of social media content should be within this range.
- `segment_size`: Maximum number of items to include in a single file. `null` means that all items should be placed in a single file.

## Export status message

Export status messages allow an exporter to provide information on the exports it performs. Example:
• The routing key will be `export.status.<social media platform>.<type>`. For example, `export.status.flickr.flickr_user`.

• `status`: Valid values are `running`, `completed success` or `completed failure`.

• `infos`, `warnings`, and `errors`: Lists of messages. A message should be an object (i.e., dictionary) containing a `code` and `message` entry. Codes should be consistent to allow message consumers to identify types of messages.

• `service`, `host`, and `instance` identify what performed the harvest. `service` is the name of the harvester. `host` is an identifier for the location of the harvest, e.g., the Docker container id. `instance` is an identifier for the process of the service on the host, e.g., the PID. The is helps in cases there may be multiple instances of a service on a host.
• genindex
• modindex
• search

21.1 Funding history

• Development of this project was supported by a grant (#NARDI-14-50017-14) from the National Historical Publications & Records Commission to George Washington University Libraries from 2014-2017.

• Development of the Sina Weibo harvester was supported by a grant from the Council on East Asian Libraries.

• Prior development of SFM under the previous repository was supported by a grant (#LG-46-13-0257-13) from the Institute of Museum and Library Services to George Washington University Libraries from 2013-2014.