Social Feed Manager:
A Social Media Capture Tool

Digital Preservation Coalition
Web Archiving & Preservation Working Group - General Meeting
December 9, 2021

Dan Kerchner
George Washington University Libraries
Washington, DC, USA
kerchner@gwu.edu
@DanKerchner
Agenda

● About SFM
● Collecting with SFM: A Walkthrough
● SFM workflows
About
Social Feed Manager ("SFM")
Social Feed Manager software

- Open source software by GW Libraries.
- User interface for collecting, managing, and exporting social media data.
- Collect from Twitter, Tumblr, Flickr, Sina Weibo.
- Libraries run this for their users as a service. (Not typically a local install on your laptop.)

More: go.gwu.edu/sfm
Institutions using SFM include

The George Washington University
Washington, DC

Stanford University

Virginia Tech

Penn State

Radcliffe Institute for Advanced Study
Harvard University
Social media capture via API, versus web capture

- Web capture captures "a" user experience (content + presentation)
- [Social media] API capture gives you:
  - Data you can't get by scraping because it's not rendered
  - Data in a structured format, easier for analyzing (usually JSON)
  - Subsets of interest that you might not be able to get via web capture
- Units of capture: {site, page, ...} vs. {account, post, ...}
- API capture usually requires credentials
Congratulations to CCAS alums Sally Nuamah, BA '11, and Tara Dorfman, BA '11, who were both named to Forbes 2019 30 Under 30 lists! #GWCCAS #GWU
Congratulations to CCAS alums Sally Nuamah, BA '11, and Tara Dorfman named to Forbes 2019 30 Under 30 lists! #GWCCAS #GWU https://t.co/kPWyKrJ9UX
Social media capture via API using SFM

- SFM uses the platforms' free standard APIs* (no scraping)
- API responses are JSON
- For Twitter capture, SFM uses **twarc**
- Collected content is stored on the file server as **WARC files**
- Does not currently capture media in posts**
- SFM handles rate limiting, authentication, organizing collections

*We are working on Twitter v2 API (incl. Academic Researcher)
**Used to with Heretrix; can access & download via links in posts
Social Feed Manager: Archivable data formats

Elemental unit of interest

- Social media post

Data with provenance:

- WARC (HTTP response w/JSON body)

Derivatives:

- JSON, CSV, Excel, identifier list

Which should we preserve?
<table>
<thead>
<tr>
<th>id</th>
<th>tweet_url</th>
<th>created_at</th>
<th>parsed_cre</th>
<th>user_scre text</th>
<th>tweet_typ</th>
<th>coordin</th>
<th>hashtags</th>
<th>media</th>
<th>urls</th>
<th>favorite_cc</th>
</tr>
</thead>
<tbody>
<tr>
<td>106970559017435</td>
<td><a href="https://twitter.com/KyrahAltman/status/106970559017435">https://twitter.com/KyrahAltman/status/106970559017435</a></td>
<td>Mon Dec 05 2018</td>
<td>12:03:51</td>
<td>a senior majoring in human services and social justice, was a original</td>
<td></td>
<td></td>
<td>CCASOutFr</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>106967182005249</td>
<td><a href="https://twitter.com/KyrahAltman/status/106967182005249">https://twitter.com/KyrahAltman/status/106967182005249</a></td>
<td>Mon Dec 05 2018</td>
<td>12:03:51</td>
<td>a professor Maria del Carmen Montoya, featured in a rec retweet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>106934420665575</td>
<td><a href="https://twitter.com/KyrahAltman/status/106934420665575">https://twitter.com/KyrahAltman/status/106934420665575</a></td>
<td>Dec 02 2018</td>
<td>12:02:40</td>
<td>The new #GWCCAS International Buddy Program wants to help internation original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>10692671931821</td>
<td><a href="https://twitter.com/KyrahAltman/status/10692671931821">https://twitter.com/KyrahAltman/status/10692671931821</a></td>
<td>Dec 02 2018</td>
<td>12:02:40</td>
<td>Congratulations to CCAS alums Sally Nuamah, BA ’11, and Tara Dorfman, Br original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>10689623730842</td>
<td><a href="https://twitter.com/KyrahAltman/status/10689623730842">https://twitter.com/KyrahAltman/status/10689623730842</a></td>
<td>Dec 01 2018</td>
<td>12:01:20</td>
<td>a #GWCCAS junior and her peers got the opportunity to see Michelle Obama original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>106888619715248</td>
<td><a href="https://twitter.com/KyrahAltman/status/106888619715248">https://twitter.com/KyrahAltman/status/106888619715248</a></td>
<td>Dec 01 2018</td>
<td>12:01:20</td>
<td>Last week, @CorcoranGW hosted Washington D.C.’s annual National Portfolio Day original</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>in_reply_to_id</th>
<th>in_reply_to_tweet_id</th>
<th>in_reply_to_lang</th>
<th>place</th>
<th>possibly_reply</th>
<th>retweet_count</th>
<th>retweet_origin</th>
<th>user_id</th>
<th>user_create</th>
<th>user_default</th>
<th>user_desc</th>
<th>user_favor</th>
<th>user_follow</th>
<th>user_friend</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>FALSE</td>
<td>0</td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>TRUE</td>
<td>112696660748309</td>
<td>CorcoranGW</td>
<td>216058845</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SFM Workflows

Twitter → Social Feed Manager → GW Tweetsets → Harvard Dataverse

Librarians, Students, Faculty, Researchers, within GW → GW Intranet

Full Twitter Data as JSON or CSV/Excel derivatives

Tweet IDs only

Students/Faculty/Researchers within GW → Public Internet

Public

Students/Faculty/Researchers within GW

Public

credit: developer icons by Oksana Latysheva from the Noun Project
SFM Walkthrough
Social Feed Manager empowers researchers and archivists to build collections of social media data from multiple platforms.

Log In

If you have not created an account yet, then please sign up first.

Username

Password

Forgot your password?

☐ Remember me

Log In

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

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

In addition to respecting the platforms’ terms, as a user of Social Feed Manager and data collected within it, it is your responsibility to consider the ethical aspects of collecting and using social media data. Your discipline or professional organization may offer guidance. Here are some ethical and privacy guidelines you may want to consider.
Credentials

Credentials are used to authorize Social Feed Manager to collect data from Twitter. Authorize Social Feed Manager by connecting your account or adding credentials.

Twitter Credential  Flickr Credential  Weibo Credential  Tumblr Credential

Name  Date Added

GWKerchner's twitter credential  Oct. 12, 2018, 10:31:34 a.m. EDT

Connect Twitter Account  Add Twitter Credential

Authorize Social Feed Manager (GW Sandbox) to use your account?

Username or email
Password

Remember me  Forgot password?

Sign In  Cancel

This application will be able to:
- Read Tweets from your timeline.
- See who you follow.

Will not be able to:
- Follow new people.
- Update your profile.
- Post Tweets for you.
- Access your direct messages.
- See your email address.
- See your Twitter password.

By GW Libraries
go.gwu.edu/sfm
Social Feed Manager, GW Sandbox
Add Twitter user timeline

* indicates a required field.

**Collection name***
GWU official accounts

**Description**

**Public link**
Link to a public version of this collection, e.g., in a data repository.

**Credential***
GWKerchner's twitter credential

**Schedule***

- Incremental harvest
  - Only collect new items since the last data retrieval.
- Automatically delete seeds for deleted / not found accounts.
- Automatically delete seeds for suspended accounts.
- Automatically delete seeds for protected accounts.

**End date**

If blank, will continue until stopped.

**Sharing***

- Group only

Who else can view and export from this collection. Select "All other users" to share with all Social Feed Manager users.

**Change Note**

Further information about this addition.

[Save] [Cancel]
Request Export

Seed choice*
- All seeds
- Active seeds only
- Selected seeds only

Export format*
- Excel (XLSX)

Maximum number of items per file
- 250,000

Deduplicate (remove duplicate posts)

Limit by item date range
Item date start
Item date end

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

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

In addition to respecting the platforms’ terms, as a user of Social Feed Manager and data collected within it, it is your responsibility to consider the ethical aspects of collecting and using social media data. Your discipline or professional organization may offer guidance. Here are some ethical and privacy guidelines you may want to consider.
GWU official accounts

Twitter user timeline

New collection added. You can now add seeds.

At least 1 active seed must be added before harvesting can be turned on.

Data collected: 0 files (0 bytes)

Seeds

<table>
<thead>
<tr>
<th>Active</th>
<th>Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Link</th>
<th>Twitter accounts</th>
<th>User ID</th>
<th>Messages</th>
</tr>
</thead>
</table>

Add Seed  Bulk Add Seeds
Add Twitter user timeline seeds

Seeds type*

- Screen Name
- User id

Bulk Seeds*

- gw_sports
- GWmedia
- GWtweets
- gwucolumnian

Enter each seed on a separate line.

Change Note

Further information about this addition.

Save  Cancel
### GWU official accounts

**Twitter user timeline**
Collection is active. Turn off to edit.

Next harvest at Dec. 4, 2018, 5:29:18 p.m. EST

**Data collected:** 1 file (4.7 MB)

**Stats:**
- tweets: 12,865

### Seeds

<table>
<thead>
<tr>
<th>Link</th>
<th>Twitter accounts</th>
<th>User ID</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>🔄</td>
<td>GWmedia</td>
<td>59863790</td>
<td></td>
</tr>
<tr>
<td>🔄</td>
<td>gw_sports</td>
<td>54627173</td>
<td></td>
</tr>
<tr>
<td>🔄</td>
<td>GWtweets</td>
<td>28101965</td>
<td></td>
</tr>
<tr>
<td>🔄</td>
<td>gwucolumnbian</td>
<td>54639174</td>
<td></td>
</tr>
</tbody>
</table>
Export files for GWU official accounts

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>6b0da4c9bb0e4087a48c8793effe0bab-README.txt</td>
<td>2.0 KB</td>
</tr>
<tr>
<td>6b0da4c9bb0e4087a48c8793effe0bab_001.xlsx</td>
<td>2.8 MB</td>
</tr>
</tbody>
</table>

See the data dictionary for more information about the fields in the export.

See the guidance on citing SFM and datasets.

**Status:** Success

**Format:** xlsx

**Selected seeds:** All seeds

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

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

In addition to respecting the platforms' terms, as a user of Social Feed Manager and data collected within it, it is your responsibility to consider the ethical aspects of collecting and using social media data. Your discipline or professional organization may offer guidance. Here are some ethical and privacy guidelines you may want to consider.
Collection Sets

A collection set is a group of collections around a particular topic or theme. Collections sets are active when there is at least one active collection within them. Collection sets are inactive when all collections have been deactivated and are no longer harvesting.

<table>
<thead>
<tr>
<th>Name</th>
<th>Collections</th>
<th>Date Added</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>115th U.S. Congress</td>
<td>3 collections</td>
<td>Jan. 27, 2017, 10:47:12 a.m. EST</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>Alaska Earthquake</td>
<td>2 collections</td>
<td>Nov. 30, 2018, 1:37:25 p.m. EST</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>China Anti-Corruption</td>
<td>48 collections</td>
<td>June 29, 2016, 11:49:14 p.m. EDT</td>
<td>CEAL Grant</td>
</tr>
<tr>
<td>Climate change</td>
<td>1 collection</td>
<td>Sept. 21, 2017, 8:11:54 a.m. EDT</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>Corcoran</td>
<td>1 collection</td>
<td>March 29, 2017, 5:17:54 p.m. EDT</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>Foreign leaders</td>
<td>1 collection</td>
<td>March 11, 2018, 9:43:54 p.m. EDT</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>Governors</td>
<td>1 collection</td>
<td>March 11, 2018, 9:34:03 p.m. EDT</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
</tbody>
</table>
SFM Workflows

Twitter → Social Feed Manager → GW Tweetsets → Harvard Dataverse

Librarians, Students, Faculty, Researchers, within GW

GW Intranet → Public Internet

Students, Faculty, Researchers, within GW

Public

Full Twitter Data as JSON or CSV/Excel derivatives

Tweet IDs only

Students/Faculty/Researchers within GW

Public

credit: developer icons by Oksana Latysheva from the Noun Project
SFM software

- Open source on GitHub
- Issues tracked in the sfm-ui repository: we welcome requests!
- Documentation for installation and use: sfm.readthedocs.io
- Main project site (blog posts and more): go.gwu.edu/sfm
Thank you

Social Feed Manager
sfm@gwu.edu
go.gwu.edu/sfm

Dan Kerchner
kerchner@gwu.edu
@DanKerchner