

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

VT
VIRGINIA TECHTM



PennState



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



GW Arts & Sciences

@gwucolumbian

Following



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](#)



11:29 AM - 2 Dec 2018

3 Retweets 9 Likes



↻ 3

♡ 9



```
{
  created_at: "Sun Dec 02 16:29:00 +0000 2018",
  id: 1069267199318216700,
  id_str: "1069267199318216704",
  full_text: "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",
  truncated: false,
+ display_text_range: [...],
- entities: {
  - hashtags: [
    - {
      text: "GWCCAS",
      - indices: [
        132,
        139
      ]
    },
    - {
      text: "GWU",
      - indices: [
        140,
        144
      ]
    }
  ],
  symbols: [ ],
  user_mentions: [ ],
  urls: [ ],
- media: [
  - {
    id: 1069267196357091300,
    id_str: "1069267196357091328",
    - indices: [
      145,
      168
    ],
    media_url: "http://pbs.twimg.com/media/DtbM2ZDXoAAMvO7.jpg",
    media_url_https: "https://pbs.twimg.com/media/DtbM2ZDXoAAMvO7.jpg",
    url: "https://t.co/kPWYKrJ9Ux",
    display_url: "pic.twitter.com/kPWYKrJ9Ux",
    expanded_url: "https://twitter.com/gwucolumbian/status/10692671993182
type: "photo".
```

```
+ - view source
in_reply_to_status_id: null,
in_reply_to_status_id_str: null,
in_reply_to_user_id: null,
in_reply_to_user_id_str: null,
in_reply_to_screen_name: null,
- user: {
  id: 54639174,
  id_str: "54639174",
  name: "GW Arts & Sciences",
  screen_name: "gwucolumbian",
  location: "Washington, D.C.",
  description: "Official Twitter account for the Columbian College of Arts and
Sciences at The George Washington University. Home of the Engaged Liberal
Arts.",
  url: "https://t.co/g0ys0T59mJ",
- entities: {
  - url: {
    - urls: [
      - {
        url: "https://t.co/g0ys0T59mJ",
        expanded_url: "http://columbian.gwu.edu",
        display_url: "columbian.gwu.edu",
        - indices: [
          0,
          23
        ]
      }
    ]
  },
  - description: {
    urls: [ ]
  }
},
protected: false,
followers_count: 4322,
friends_count: 608,
listed_count: 136,
created_at: "Tue Jul 07 18:49:10 +0000 2009",
favourites_count: 4202,
utc_offset: null,
time_zone: null,
geo_enabled: true,
verified: false,
statuses_count: 9273,
lang: "en",
```


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:

- WARCs (HTTP response w/JSON body)

Derivatives:

- JSON, CSV, Excel, identifier list

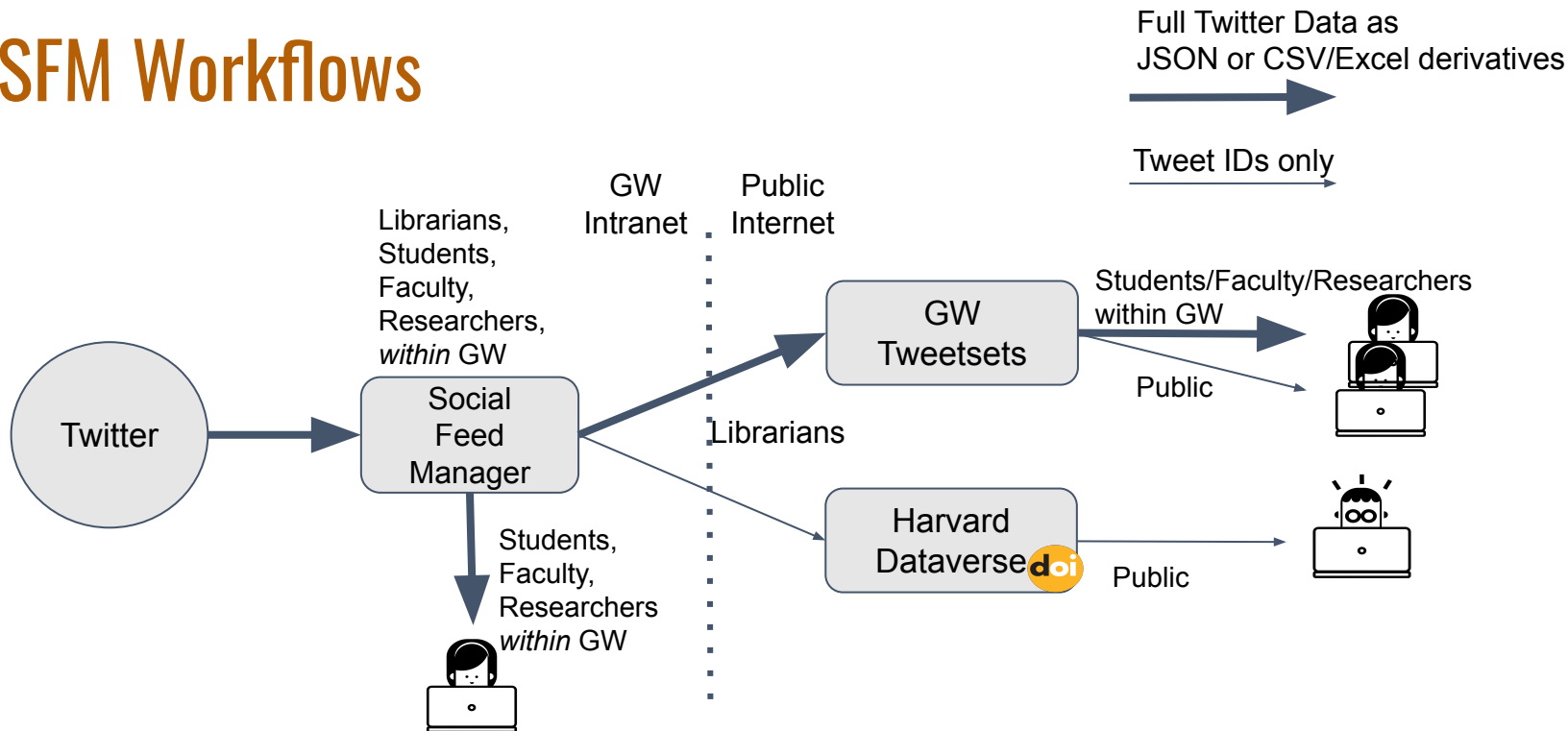
Which should we preserve?

CSV/Excel extract

id	tweet_url	created_at	parsed_created_at	user_screen_name	text	tweet_type	coordinates	hashtags	media	urls	favorite_count
106970559017435	https://twitter.com/gwucolumb/status/106970559017435	Mon Dec 04 18:05:24 -0800	2018-12-04T18:05:24-08:00	gwucolumb	Kyrarh Altman, a senior majoring in human services and social justice, was a	original		CCASOutFr		https://twitter.com/gv	0
106967441805249	https://twitter.com/gwucolumb/status/106967441805249	Mon Dec 04 18:05:24 -0800	2018-12-04T18:05:24-08:00	gwucolumb	RT @CorcoranGW: Professor Maria del Carmen Montoya, featured in a rec	retweet					0
106934420665575	https://twitter.com/gwucolumb/status/106934420665575	Sun Dec 02 18:05:24 -0800	2018-12-02T18:05:24-08:00	gwucolumb	The new #GWCCAS International Buddy Program wants to help internation	original		GWCCAS G		https://twi https://bit.	1
106926719931821	https://twitter.com/gwucolumb/status/106926719931821	Sun Dec 02 18:05:24 -0800	2018-12-02T18:05:24-08:00	gwucolumb	Congratulations to CCAS alums Sally Nuamah, BA '11, and Tara Dorfman, B	original		GWCCAS G		https://twitter.com/gv	9
106896923730842	https://twitter.com/gwucolumb/status/106896923730842	Sat Dec 01 18:05:24 -0800	2018-12-01T18:05:24-08:00	gwucolumb	A #GWCCAS junior and her peers got the opportunity to see Michelle Obam	original		GWCCAS O		https://twi https://bit.	3
106888619715248	https://twitter.com/gwucolumb/status/106888619715248	Sat Dec 01 18:05:24 -0800	2018-12-01T18:05:24-08:00	gwucolumb	Last week, @CorcoranGW hosted Washington D.C.'s annual National Portf	original		GWU GWC		https://twi https://bit.	3

in_reply_to_tweet_id	in_reply_to_user_id	in_reply_to_screen_name	lang	place	possibly_sensitive	retweet_count	retweet_or_quote_count	retweet_or_quote_tweet_id	retweet_or_quote_user_id	source	user_id	user_created_at	user_default_profile_image	user_description	user_follower_count	user_following_count	user_friends_count
			en		FALSE	0				1069666074830 CorcoranGW	54639174	Tue Jul 07 18:05:24 -0800	FALSE	Official Twi	4202	4322	608
			en			1	1069666074830	CorcoranGW	216058845	1069666074830 CorcoranGW	54639174	Tue Jul 07 18:05:24 -0800	FALSE	Official Twi	4202	4322	608
			en		FALSE	0				1069666074830 CorcoranGW	54639174	Tue Jul 07 18:05:24 -0800	FALSE	Official Twi	4202	4322	608
			en		FALSE	3				1069666074830 CorcoranGW	54639174	Tue Jul 07 18:05:24 -0800	FALSE	Official Twi	4202	4322	608
			en		FALSE	0				1069666074830 CorcoranGW	54639174	Tue Jul 07 18:05:24 -0800	FALSE	Official Twi	4202	4322	608
			en		FALSE	1				1069666074830 CorcoranGW	54639174	Tue Jul 07 18:05:24 -0800	FALSE	Official Twi	4202	4322	608

SFM Workflows



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

Social Feed Manager empowers researchers and archivists to create collections of social media data from Twitter, Tumblr, Flickr, and Sina Weibo. Read more about Social Feed Manager [here](#).

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

[Sign up for Twitter >](#)

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



Social Feed Manager (GW Sandbox)

By GW Libraries

go.gwu.edu/sfm

Social Feed Manager, GW Sandbox

☐ 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.

[Collection Sets](#) / GW Twitter Accounts

GW Twitter Accounts

 Edit

Data collected: 0 files (0 bytes)

Details ▾

Collections

No collections yet.

Add Collection ▾ 

Add Twitter user timeline

Tweets from specific accounts

Add Twitter search

Recent tweets matching a query

Add Twitter filter

Tweets in real time matching filter criteria

Add Twitter sample

A subset of all tweets in real time

Add Tumblr blog posts

Blog posts from specific blogs

Add Flickr user

Posts and photos from specific accounts

Add Weibo timeline

Posts from a user and the user's friends

	User
4, 2018, 7:01:42 p.m. EDT	system

[SFM UI 2.0.2](#) | [User Guide](#) | [Documentation](#) |

[Collection Sets](#) / [GW Twitter Accounts](#) / Add Twitter user timeline

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

☒ 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.

Schedule*

Every week

How frequently you want data to be retrieved.

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

[Collection Sets](#) / [GW Twitter Accounts](#) / [GWU official accounts](#) / [Request Export](#)

Request Export

Seed choice*

- ☒ All seeds
- ☐ Active seeds only
- ☐ Selected seeds only

None selected ▾

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

 × 

The timezone for dates entered here are America/New_York. Adjustments will be made to match the time zone of the items. For example, dates in tweets are UTC.

Limit by harvest date range

Harvest date start


 × 

Limit by harvest date range

Harvest date start

 × 

Harvest date end

 × 

Export

Cancel

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](#) / [GW Twitter Accounts](#) / GWU official accounts

GWU official accounts



Twitter user timeline

⏻ Turn on

⏻ Deactivate

✎ Edit

📄 Export

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)

Details

Seeds

📄 Download seed list

Active

Deleted

Search

Link

Twitter accounts

User ID

Messages

Add Seed

Bulk Add Seeds

Add Twitter user timeline seeds

Seeds type*

☒ Screen Name

☐ User id

Bulk Seeds*

gw_sports
GWmedia
GWtweets
gwucolumbian

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.

Turn off

Deactivate

Edit

Export

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

Data collected: 1 file (4.7 MB)

Stats:

- tweets: 12,865

Details

Seeds

Download seed list

Active 4 Deleted

Search

Link	Twitter accounts	User ID	Messages
	GWmedia	59863790	
	gw_sports	54627173	
	GWtweets	28101965	
	gwucolumbian	54639174	

[Collection Sets](#) / [GW Twitter Accounts](#) / [GWU official accounts](#) / [Export](#)

Export files for GWU official accounts

Filename	Size
6b0da4c9bb0e4087a48c8793effe0bab-README.txt	2.0 KB
6b0da4c9bb0e4087a48c8793effe0bab_001.xlsx	2.8 MB

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

[Details ▾](#)

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.

Active 22

Inactive 13

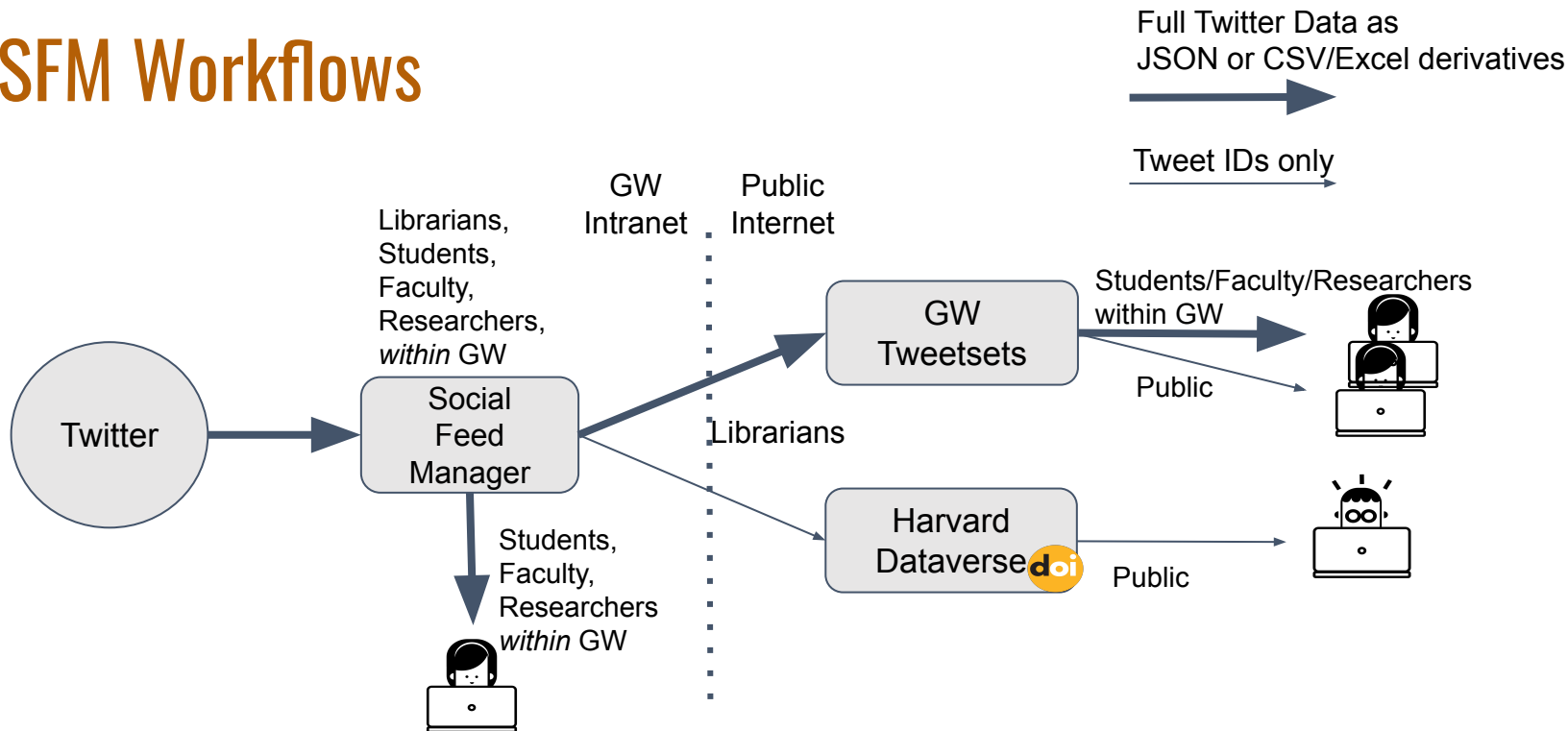
Shared 23

Other Active 109

Other Inactive 12

Name	Collections	Date Added	Groups
115th U.S. Congress	3 collections	Jan. 27, 2017, 10:47:12 a.m. EST	GW Libraries Scholarly Technology Group
2017-2020 Federal Term	2 collections	Jan. 20, 2017, 11:26:54 a.m. EST	GW Libraries Scholarly Technology Group
Alaska Earthquake	2 collections	Nov. 30, 2018, 1:37:25 p.m. EST	GW Libraries Scholarly Technology Group
China Anti-Corruption	48 collections	June 29, 2016, 11:49:14 p.m. EDT	CEAL Grant
Climate change	1 collection	Sept. 21, 2017, 8:11:54 a.m. EDT	GW Libraries Scholarly Technology Group
Corcoran	1 collection	March 29, 2017, 5:17:54 p.m. EDT	GW Libraries Scholarly Technology Group
Foreign leaders	1 collection	March 11, 2018, 9:43:54 p.m. EDT	GW Libraries Scholarly Technology Group
Governors	1 collection	March 11, 2018, 9:34:03 p.m. EDT	GW Libraries Scholarly Technology Group

SFM Workflows



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