Social Feed Manager

Digital Preservation Coalition
Web Archiving & Preservation Webinar Series
February 6, 2020

Dan Kerchner
kerchner@gwu.edu
@DanKerchner

Laura Wrubel
lwrubel@gwu.edu
@liblaura

Agenda

- Project background
- Current uses
- Collecting from Twitter
- SFM walkthrough
- Ethical and technical considerations

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

Project Background

- Initial development: 2013
- Version 1.0: 2016
- Grant funding:
  - IMLS
  - National Historical Publications & Records Commission (NARA)
  - CEAL

Institutions using SFM include

Social media research

INFORMATION, COMMUNICATION & SOCIETY, 2017
VOL. 20, NO. 9, 1330–1346
https://doi.org/10.1080/1369118X.2017.1328521

Routledge
Taylor & Francis Group

Check for updates

Populist communication by digital means: presidential Twitter in Latin America

Silvio Waisbord[a] and Adriana Amado[b]

[a]School of Media and Public Affairs, George Washington University, Washington, DC, USA; [b]Universidad de La Matanza, San Justo, Argentina

ABSTRACT
In this paper, we analyze the uses of Twitter by populist presidents in contemporary Latin America in the context

ARTICLE HISTORY
Received 30 November 2016

Check for updates

Research Article

Twitter Makes It Worse: Political Journalists, Gendered Echo Chambers, and the Amplification of Gender Bias

The International Journal of Press/Politics
2018, Vol. 23(3) 324–344

DOI: 10.1177/1940161218781254
journals.sagepub.com/home/hij
SAGE

Nikki Usher[1], Jesse Holcomb[2], and Justin Littman[3]

Abstract
Given both the historical
nalism and
nication, th
der bias a
characteriz
p-peer rela
urnalists fo
inequities
e peers al
other. Th
of gender
rch but als
olitical cor

THE GEORGE WASHINGTON UNIVERSITY
THE ELLIOTT SCHOOL OF INTERNATIONAL AFFAIRS

THE INSTITUTE FOR MIDDLE EAST STUDIES

IMES CAPSTONE PAPER SERIES

Overcoming Failure: How Extremist Organizations Leverage Setbacks with Powerful Propaganda

Rachel N. Ridley
Mattisan S. Rowan

May 2017

HE INSTITUTE FOR MIDDLE EAST STUDIES
LLIOTT SCHOOL OF INTERNATIONAL AFFAIRS
THE GEORGE WASHINGTON UNIVERSITY

Political Communication, 36:195–202, 2019

ISSN: 1058-4609 print / 1091-7675 online
DOI: https://doi.org/10.1080/10584609.2018.1548412

Routledge
Taylor & Francis Group

Check for updates

The Forum

Who's Elite and How the Answer Matters to Politics

BABAK BAHADOR, ROBERT ENTMAN, and CURD KNÜPFER

SOUTH ASIAN REVIEW
https://doi.org/10.1080/02759527.2019.1599553

Routledge
Taylor & Francis Group

Check for updates

Inside and Outside the Literary Marketplace: The Digital Products of Amitav Ghosh, Arundhati Roy, and Salman Rushdie

Tawnya Azar

English Department, George Mason University, Fairfax, VA, USA

ABSTRACT
For Indian authors writing in English, the changing landscape of publishing in the digital age presents both unique challenges and opportunities for claiming their authority as public intellectuals and producing content that exists both within the established literary marketplace and outside of it. This paper examines the digital products and platforms of three such authors, Salman Rushdie, Amitav Ghosh, and Arundhati Roy, all of whom can claim

ARTICLE HISTORY
Received 10 July 2018
Accepted 11 December 2018

KEYWORDS
Digital; humanities; literary; marketplace; Salman Rushdie; Amitav Ghosh; Arundhati Roy

ism, gende

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/DtbM2ZDXoAAMvQ7.jpg",
        media_url_https: "https://pbs.twimg.com/media/DtbM2ZDXoAAMvQ7.jpg",
        url: "https://t.co/kPWyKrJ9Ux",
        display_url: "pic.twitter.com/kPWyKrJ9Ux",
        expanded_url: "https://twitter.com/gwucolumbian/status/10692671993182
        type: "photo",
```

```
  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 an
    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",
```

Collecting from Twitter's APIs

- SFM uses the free standard APIs
- API responses are JSON
- SFM handles rate limiting, authentication, organizing collections
- Requires one set of Twitter developer keys for the application and Twitter authentication for each user collecting data.

Twitter collection types

- User timeline: up to 3,200 tweets in the past, for each account
- Search: sample of tweets from the past 7-10 days
- Filter stream: real-time, forward-looking only
- Sample stream ~ 0.5% sample of all tweets; ~2GB per day

Collecting from Twitter: Data Extracts

- JSON
- CSV and Excel (selected fields from JSON)
- Filtered by date range, screen name, etc.

CSV/Excel extract

id	tweet_url	created_at	parsed_cre	user_scree	text	tweet_typ	coordin	hashtags	media	urls	favorite_cc
106970559017435	https://twitte	Mon Dec 0	2018-12-03	gwucolumb	Kyrah Altman, a senior majoring in human services and social justice, was a	original		CCASOutFr	https://twitter.com/gv		0
106967441805249	https://twitte	Mon Dec 0	2018-12-03	gwucolumb	RT @CorcoranGW: Professor Maria del Carmen Montoya, featured in a rec	retweet					0
106934420665575	https://twitte	Sun Dec 02	2018-12-02	gwucolumb	The new #GWCCAS International Buddy Program wants to help internation	original		GWCCAS G	https://twi	https://bit.	1
106926719931821	https://twitte	Sun Dec 02	2018-12-02	gwucolumb	Congratulations to CCAS alums Sally Nuamah, BA '11, and Tara Dorfman, BA	original		GWCCAS G	https://twitter.com/gv		9
106896923730842	https://twitte	Sat Dec 01	2018-12-01	gwucolumb	A #GWCCAS junior and her peers got the opportunity to see Michelle Obam	original		GWCCAS O	https://twi	https://bit.	3
106888619715248	https://twitte	Sat Dec 01	2018-12-01	gwucolumb	Last week, @CorcoranGW hosted Washington D.C.'s annual National Portfo	original		GWU GWC	https://twi	https://bit.	3

in_reply_t	in_reply_t	in_reply_t	lang	place	possibly_s	retweet_c	retweet_or_qu	retweet_or_qu	retweet_o	source	user_id	user_creat	user_defau	user_descr	user_favou	user_follo	user_frien
			en		FALSE	0				<a href="ht	54639174	Tue Jul 07 1	FALSE	Official Twi	4202	4322	608
			en			1	1069666074830	CorcoranGW	216058845	<a href="ht	54639174	Tue Jul 07 1	FALSE	Official Twi	4202	4322	608
			en		FALSE	0				<a href="ht	54639174	Tue Jul 07 1	FALSE	Official Twi	4202	4322	608
			en		FALSE	3				<a href="ht	54639174	Tue Jul 07 1	FALSE	Official Twi	4202	4322	608
			en		FALSE	0				<a href="ht	54639174	Tue Jul 07 1	FALSE	Official Twi	4202	4322	608
			en		FALSE	1				<a href="ht	54639174	Tue Jul 07 1	FALSE	Official Twi	4202	4322	608

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

Username

Password

Password

Forgot your password?

Remember me

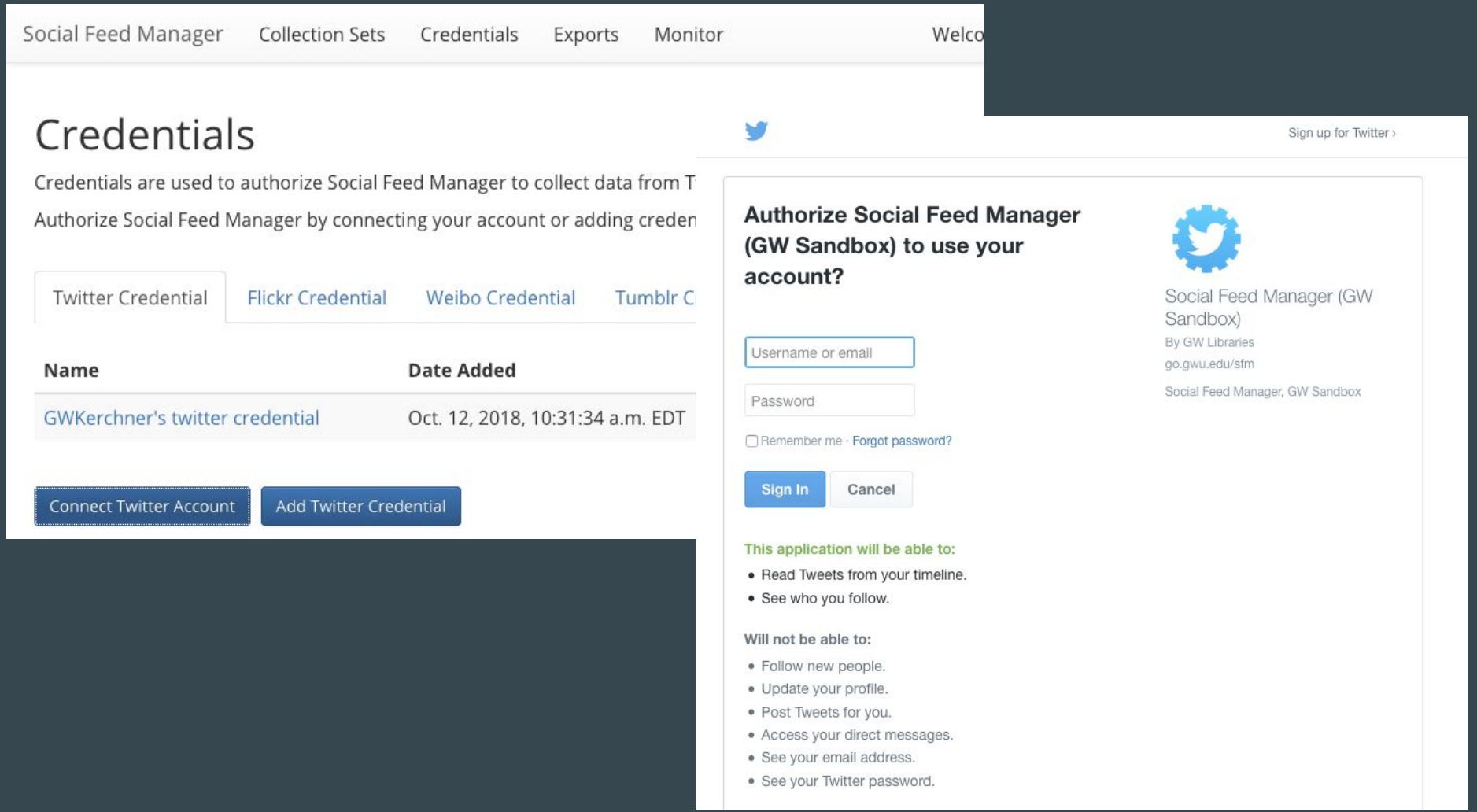
Log in

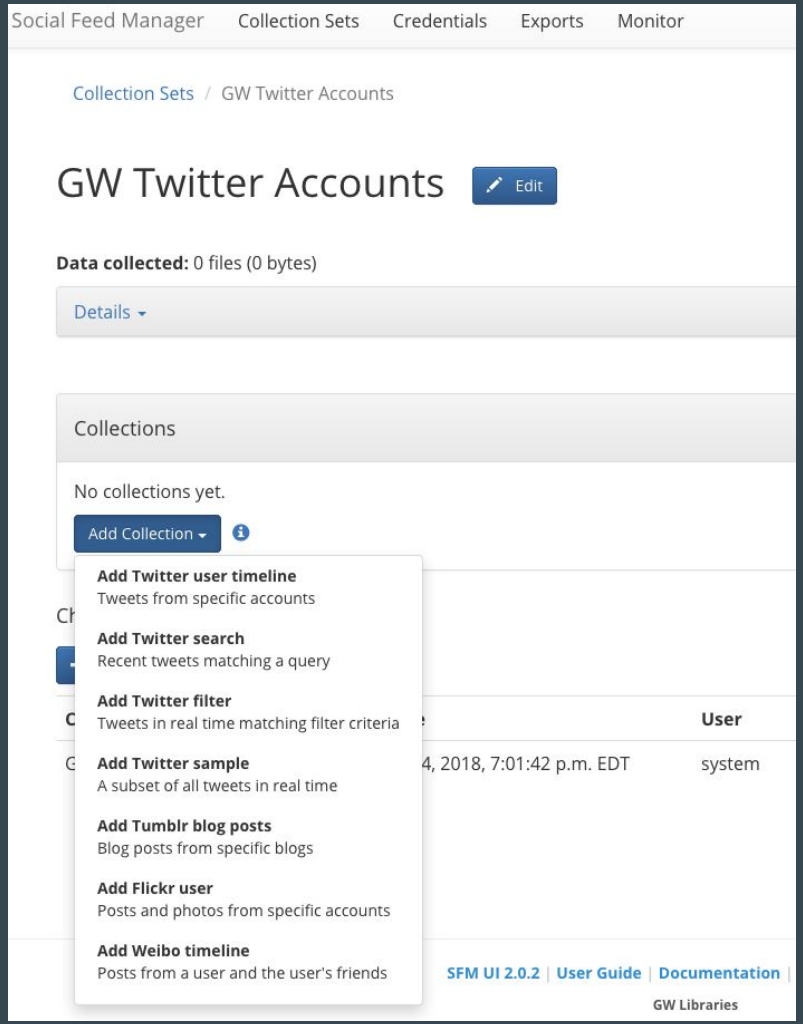
Social Feed Manager empowers researchers and archivists to create collections of social media data from Twitter, Tumblr, Flickr, and Sina Weibo. Read more about Social Feed Manager here.

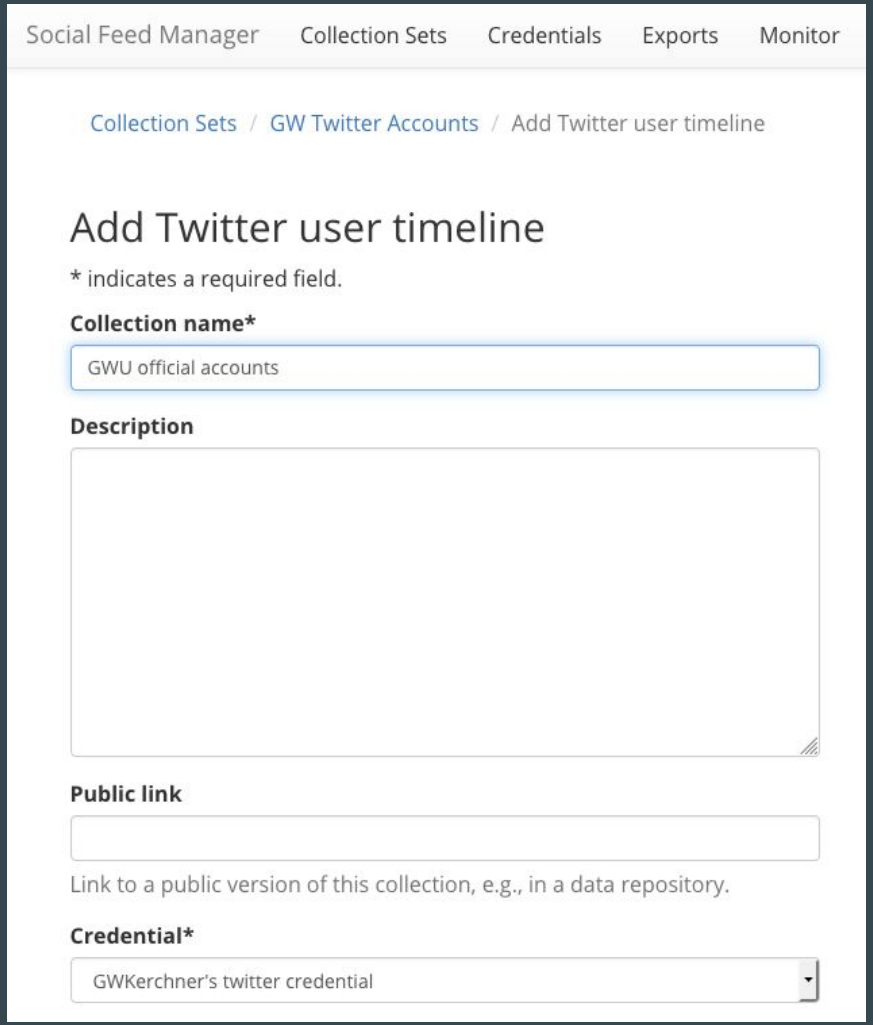
Collecting and using data from social media platforms is subject to those platforms' terms (Twitter, Flickr, Sina Weibo, Tumblr), as you agreed to them when you created your social media account. Social Feed Manager respects those platforms' terms as an application (Twitter, Flickr, Sina Weibo, Tumblr).

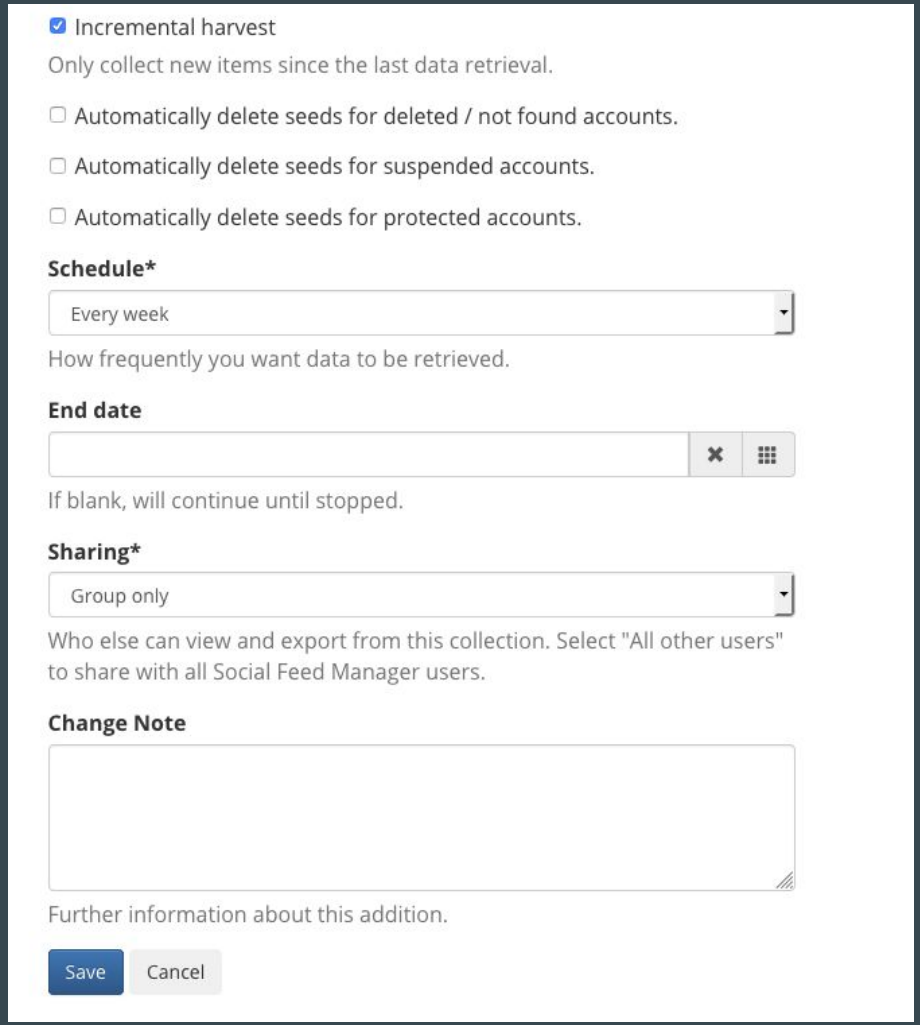
Social Feed Manager provides data to you for your research and academic use. Social media platforms' terms of service **generally do not allow republishing of full datasets,** and you should refer to their terms to understand what you may share. Authors typically retain rights and ownership to their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

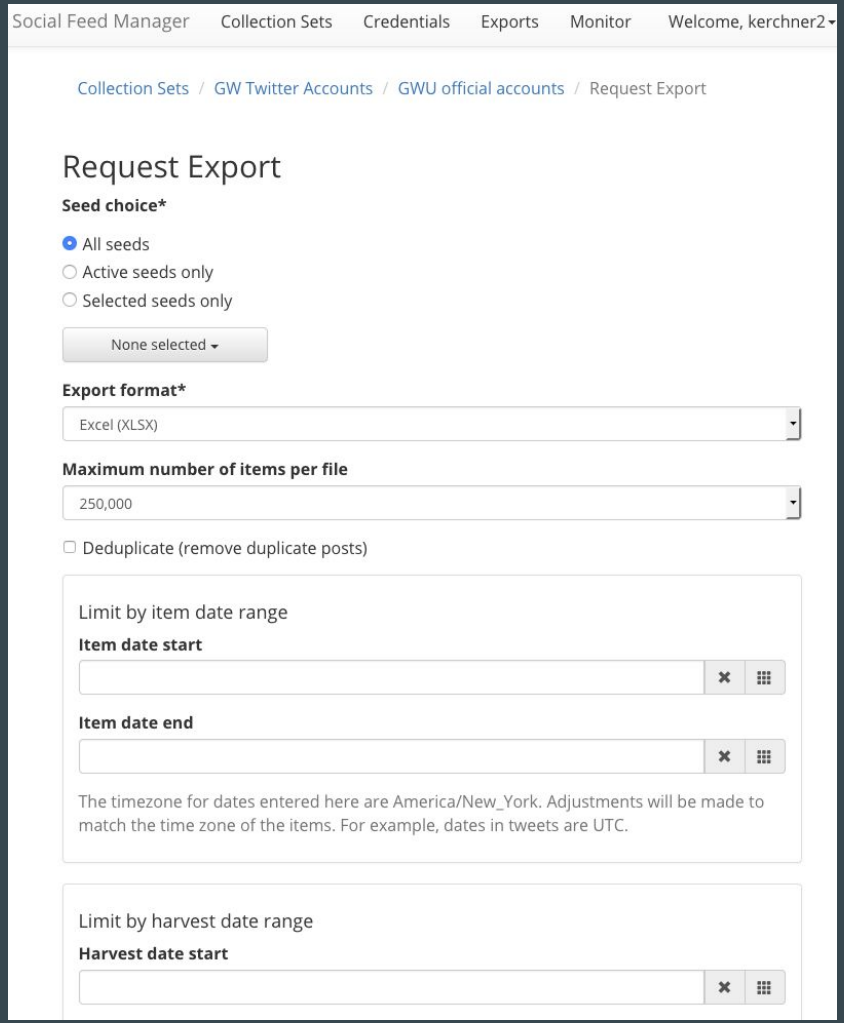
In addition to respecting the platforms' terms, as a user of Social Feed Manager and data collected within it, it is your responsibility to consider the ethical aspects of collecting and using social media data. Your discipline or professional organization may offer guidance. Here are some ethical and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

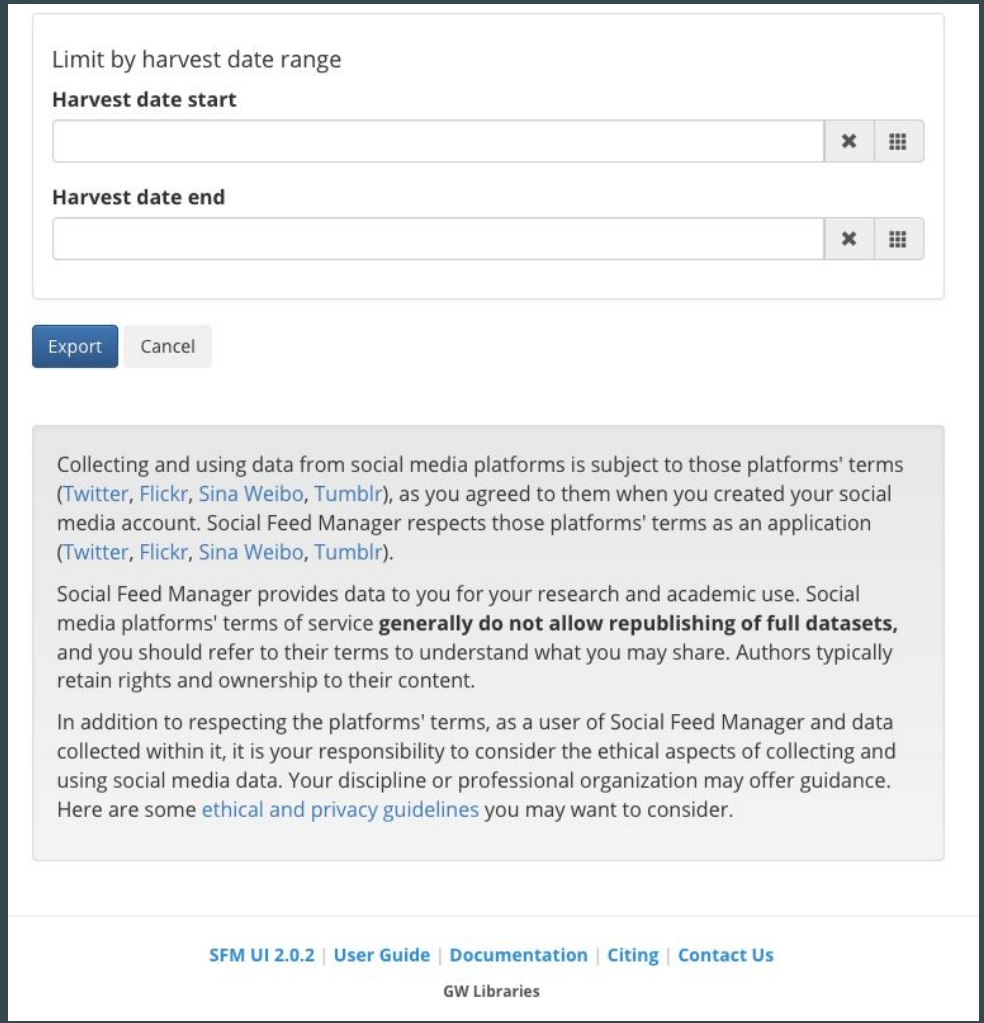

Social Feed Manager
Collection Sets
Credentials
Exports
Monitor
Welco
Credentials
Credentials are used to authorize Social Feed Manager to collect data from T
Authorize Social Feed Manager by connecting your account or adding creden
Twitter Credential
Flickr Credential
Weibo Credential
Tumblr C
Name
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?
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.
Social Feed Manager (GW Sandbox)
By GW Libraries
go.gwu.edu/sfm
Social Feed Manager, GW Sandbox


Social Feed Manager
Collection Sets
Credentials
Exports
Monitor
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 |
GW Libraries


Social Feed Manager
Collection Sets
Credentials
Exports
Monitor
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


Social Feed Manager
Collection Sets
Credentials
Exports
Monitor
Welcome, kerchner2
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.
SFM UI 2.0.2 | User Guide | Documentation | Citing | Contact Us
GW Libraries

Social Feed Manager | Collection Sets | Credentials | Exports | Monitor | Welcome, kerchner2

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

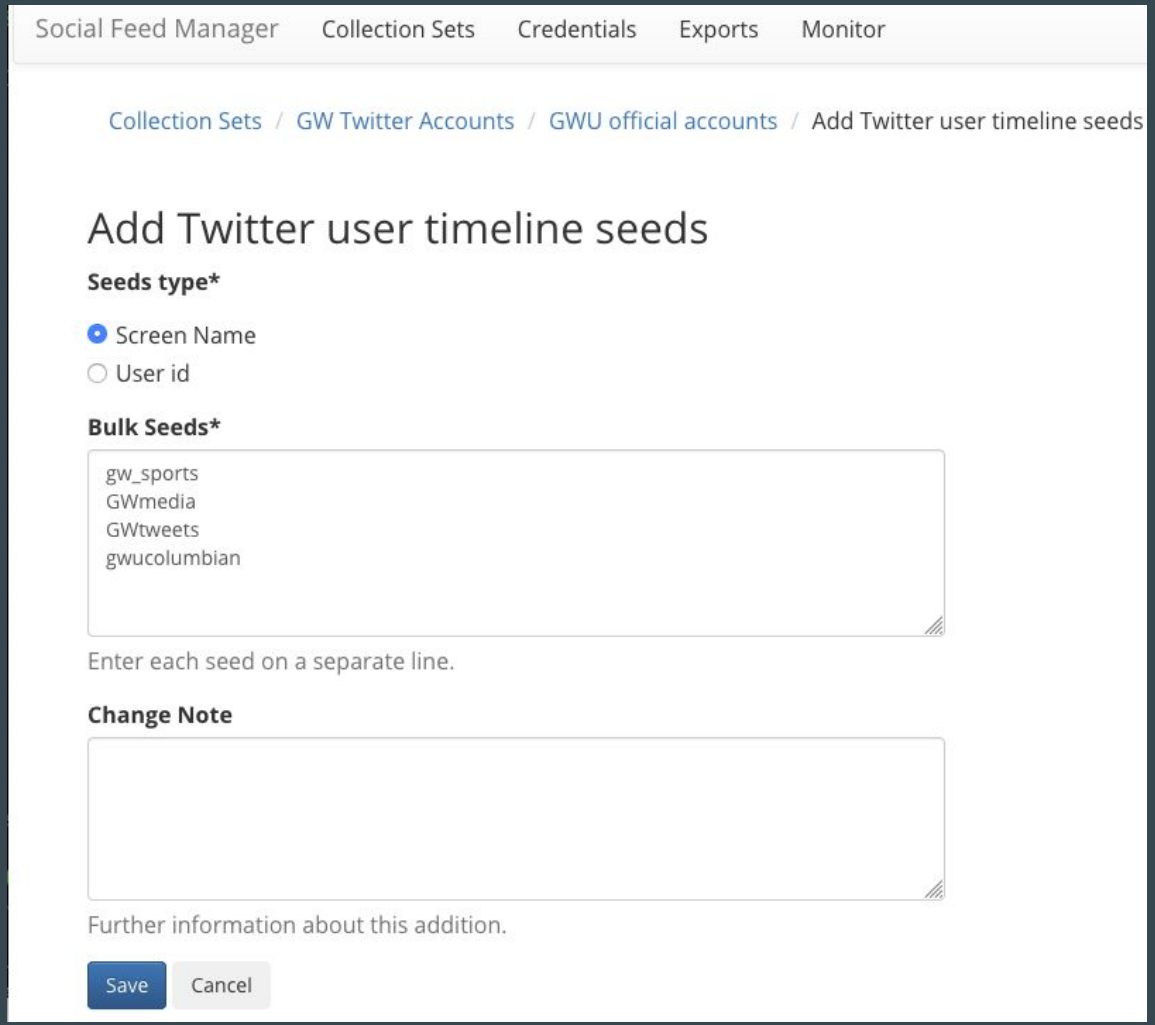
Download seed list

Active | Deleted

Search

Link	Twitter accounts	User ID	Messages

Add Seed | Bulk Add Seeds


Social Feed Manager
Collection Sets
Credentials
Exports
Monitor
Collection Sets / GW Twitter Accounts / GWU official accounts / Add Twitter user timeline 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

Collection Sets / GW Twitter Accounts / GWU official accounts

GWU official accounts

Turn off | Deactivate | Edit | Export

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

Details

Seeds

Download seed list

Active 4 | Deleted

Search

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

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

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

Technical info

Requires:

- Twitter application credentials
- Linux server
- Storage

Social Feed Manager runs as a set of Docker containers

- Application written in Python
- Small database (postgresql) for user info and metadata
- Collected content is stored on the file server as WARC files

Ethical and privacy concerns

Social media data comes from people

- Consider the impact of your work on the creator of the social media. Be thoughtful collecting social media of vulnerable individuals.
- When publishing, get permission from creator for quotes from social media if possible, do not rely on anonymizing posts.
- Respect a user's deleting their content.
- Be familiar with platform terms of use. Don't republish full datasets and share in accordance with terms (e.g., tweet ids only)

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

Laura Wrubel
lwrubel@gwu.edu
@liblaura

TweetSets
tweetsets.library.gwu.edu

Select source datasets

Previous datasets

Select the source datasets from which to create a new dataset.

Other datasets that we've collected:

- Members of the 116th Congress
- Healthcare
- Tax reform
- Hurricane Florence
- Hurricane Michael
- Solar Eclipse

Contact us if you'd like any of these loaded into TweetSets.

- [ ] 2016 United States Presidential Election: Democratic Candidate User Timelines **22,251 tweets**
- [ ] 2016 United States Presidential Election: Democratic Convention **8,340,668 tweets**
- [ ] 2016 United States Presidential Election: Democratic Party User Timelines **11,846 tweets**
- [ ] 2016 United States Presidential Election: Election Day Filter **3,626,254 tweets**
- [ ] 2016 United States Presidential Election: Election Filter **251,077,140 tweets**
- [ ] 2016 United States Presidential Election: First Presidential Debate Filter **3,182,243 tweets**
- [ ] 2016 United States Presidential Election: Republican Candidate User Timelines **52,395 tweets**
- [ ] 2016 United States Presidential Election: Republican Convention **6,248,688 tweets**