Capturing the Ephemeral: Collecting Social Media with Social Feed Manager

@bergisjules @dchud @dankerchner @liblaura
George Washington University Libraries

CNI Fall Forum - 2013-12-09 - Washington, DC

a traditional project

- save the time of the researcher

- at-risk e-resources, licensing

- expand scope of collection development

save the time
of the researcher

"How Mainstream News Outlets Use Twitter" (2011)

- GWU's Prof. Kimberly Gross and students
- Pew Research Center's Project for Excellence in Journalism
- "news agenda these organizations promoted on Twitter closely matches that of their legacy platforms"

Q: How did they collect their data?

A: By hand.

- google reader
- copy and paste
- fold, spindle, mutilate
- excel
- ...eventually, SPSS and similar tools

will take
any help
they can get

too much work
for
too little data

(only 1000s of tweets)

copy and paste
to excel
doesn't scale

(just ask her students)

over 5,000
theses and dissertations
since 2010

saving time of researchers:

strategic advantage

what researchers ask for

- specific users, keywords
- basic values: user, date, text, counts
- 10000s, not 10000000s
- delimited files to import
- historic time periods

options
for
historical data?

Twitter-licensed data providers:

DataSift
Gnip
NTT Data
Topsy*

data providers

- friendly
- not cheap*
- more than we need
- still need tools to collect, process, etc.

what can we do ourselves?

lobster traps

when Congress turned over

- 16+ accounts deleted / hidden
- combined 105,993 followers
- 14,479 tweets saved in SFM
no longer public

@GWUArchives is using Social Feed Manager to better document student life and university culture at #GWU. #cni13f

for University Archives

- practical tool
  - instant value
    - addresses collection development gap
      - document student organizations

why Student Org records

- interest from university admin
- great representation student activity
- difficult collecting area
- not in University Archives
- active social media users

#GWU on Twitter

- highly active user community
  - students, administrators, offices
- over 400 student organizations
  - greek, cultural, social, political, activist
  - exclusively on Twitter
  - no other web presence

what we've collected

- since March 2013
- tracking 329 accounts
- 216,371 tweets
  - 10,000 tweets in one month

for Special Collections

- new type of record/collection
- enhance existing collections

and beyond…

"content on social media is likely a federal record" - NARA

NARA Bulletin 2012-02: Guidance on Managing Social Media Records. October 25, 2013. http://www.archives.gov/records-mgmt/bulletins/2014/2014-02.html

moving forward

- collection development
- metadata
- links
- images
- web archiving

the software

django/python
django-social-auth
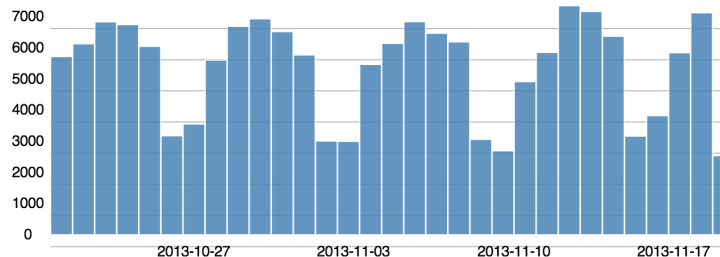tweepy

github.com/gwu-libraries/social-feed-manager

open source, MIT-style license

now with about 2,932,800 items from 1,319 users

a sampling:

JoeHeim · jamesbuck · gracenapolitano ·
PeterWhoriskey · bonniebenwick ·
AustinScottGA08 · SueMyrick ·
GWColonialArmy · emmersbrown ·
RepGaryPeters · GWCampusRec · GWToday ·
annekornblut · ajbombay · SenBillNelson ·
GWShakespeareCo · dmontyjr · lizclarketweet
· SenRockefeller · DrPhilRoe · repronkind ·
GWAthleticAlums · GWSGAC · gwupcma ·
SonsOfPitch

tweets per day

recent tweets

user	date	rt #	text
Le_Figaro	08:50:37	12	RT @quentinperinel: Au bout de 5 ans, le métier de community manager est (enfin) reconnu en France par @pole_emploi ! http://t.co/WV8qpMwM0l
Independent	08:49:54	2	Excess heat from the Northern Line to heat Islington homes http://t.co/d2RH0QPnti
Independent	08:49:31	1	£148m Euromillions winners announce marriage split a year after jackpot joy http://t.co/8Lw03ZiYJH
PravdaRu	08:49:16	0	США спасает Сирию: http://t.co/ExvxyL7aEe на @YouTube

lemondefr - 19,102 items

Twitter user 24,744,541 - created Mon Mar 16 18:44:51 +0000 2009.
88,597 tweets - 1,928,177 followers - 242 following - view in Twitter
export this data as csv for excel

date followers/following	rt #	text	view
2013-12-03 01:43:44 p.m. 1,928,177 - 242	11	Municipales : Hélène Mandroux ne se représentera pas à Montpellier http://t.co/WH5k08bfYs ↙ expand links	in twitter raw live - cached
2013-12-03 01:28:29 p.m. 1,928,177 - 242	6	Outreau : 3 000 euros d'amende requis contre Bertrand Tavernier http://t.co/bOZ3pWZ8fn ↙ expand links	in twitter raw live - cached
2013-12-03 01:11:34 p.m. 1,928,177 - 242	22	La carte mondiale de l'utilisation des sols http://t.co/kE59NcTmvm ↙ expand links	in twitter raw live - cached
2013-12-03 12:41:15 p.m. 1,928,177 - 242	21	Le Grand Paris abat ses cartes http://t.co/qTusSxGyzE ↙ expand links	in twitter raw live - cached
2013-12-03 12:04:12 p.m. 1,928,177 - 242	15	La promotion de la « mummyrexie » inquiète http://t.co/6lZMWLWCFW ↙ expand links	in twitter raw live - cached

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sfm_id	created_at	twitter_id	screen_nam	followers_co	friends_c	retweet_count	hashtags	in_reply_to_	mentions	twitter_url	is_retweet_s	is_retweet	text	url1	url1_expand	url2	url2_exp
117	3080115	2013-12-05T15:3	4.0862E+17	lemondefr	1931299	241	7	indiscrets		@lemonde_	http://twitte	TRUE	TRUE	RT @lemonde_pol: #indiscrets E	http://t.co/N	http://www.lemonde.fr/les-indisc
118	3080116	2013-12-05T15:1	4.0862E+17	lemondefr	1931299	241	79				http://twitte	FALSE	FALSE	DerniÃ¨re minute : l'ONU autori	http://t.co/s	http://lemde.fr/1g9XjWE
119	3080117	2013-12-05T15:1	4.0862E+17	lemondefr	1931299	241	12				http://twitte	FALSE	FALSE	Mobilisation en baisse de l'ense	http://t.co/u	http://lemde.fr/IRiChF
120	3080118	2013-12-05T14:2	4.086E+17	lemondefr	1931299	241	18				http://twitte	FALSE	FALSE	Le rappeur tunisien WeldÂ ElÂ 1	http://t.co/c	http://lemde.fr/IR8buz
121	3080119	2013-12-05T14:0	4.086E+17	lemondefr	1931299	241	15				http://twitte	FALSE	FALSE	Comme prÃ©vu, la BCE laisse s	http://t.co/5	http://lemde.fr/IR4R2L
122	3080120	2013-12-05T14:0	4.086E+17	lemondefr	1931299	241	7	indiscrets		@lemonde_	http://twitte	TRUE	TRUE	RT @lemonde_pol: #indiscrets E	http://t.co/2	http://www.lemonde.fr/les-indiscr
123	3078699	2013-12-05T13:1	4.0859E+17	lemondefr	1930873	241	36			@Damien_Le	http://twitte	TRUE	TRUE	RT @Damien_Leloup: Un tract d	http://t.co/K	http://imgur.com/zexAXAN
124	3078700	2013-12-05T13:0	4.0859E+17	lemondefr	1930873	241	24				http://twitte	FALSE	FALSE	Deux AlgÃ©riens dÃ©tenus Ã	http://t.co/lz	http://lemde.fr/IQRJdO
125	3078701	2013-12-05T12:4	4.0858E+17	lemondefr	1930873	241	14	Polytechnique		@LeMondeE	http://twitte	TRUE	TRUE	RT @LeMondeEduc: #Polytechn	http://t.co/8	http://ick.li/4U7NuZ
126	3078702	2013-12-05T12:3	4.0857E+17	lemondefr	1930873	241	49			@samuellau	http://twitte	TRUE	TRUE	RT @samuellaurent: La longue l	http://t.co/e	http://www.lemonde.fr/politique/
127	3078703	2013-12-05T12:2	4.0857E+17	lemondefr	1930873	241	52				http://twitte	FALSE	FALSE	Un quart des EuropÃ©ens men:	http://t.co/Y	http://lemde.fr/IQLI0A
128	3078704	2013-12-05T12:1	4.0857E+17	lemondefr	1930873	241	73				http://twitte	FALSE	FALSE	L'interminable liste des Ã©lus p	http://t.co/if	http://lemde.fr/IK2tKn
129	3078705	2013-12-05T11:1	4.0855E+17	lemondefr	1930873	241	10				http://twitte	FALSE	FALSE	La vente de La Redoute est-elle	http://t.co/if	http://lemde.fr/IQxBbl
130	3078706	2013-12-05T10:5	4.0855E+17	lemondefr	1930873	241	12			@NicolasBou	http://twitte	TRUE	TRUE	RT @NicolasBourcier BrÃ©sil : l	http://t.co/Z	http://www.lemonde.fr/ameriques
131	3078707	2013-12-05T10:4	4.0855E+17	lemondefr	1930873	241	19	indiscrets		@lemonde_	http://twitte	TRUE	TRUE	RT @lemonde_pol: #indiscrets F	http://t.co/fl	http://www.lemonde.fr/les-indiscr
132	3078708	2013-12-05T10:3	4.0855E+17	lemondefr	1930873	241	42				http://twitte	FALSE	FALSE	DÃ©mission de l'entraÃ®neur d	http://t.co/s	http://lemde.fr/IJQHQa
133	3078709	2013-12-05T10:2	4.0854E+17	lemondefr	1930873	241	14	Florange, CFDT		@CPietralun	http://twitte	TRUE	TRUE	RT @CPietralunga: Edouard Martin #Florange tÃªte de liste PS aux europÃ©enne
134	3078710	2013-12-05T10:2	4.0854E+17	lemondefr	1930873	241	42				http://twitte	FALSE	FALSE	Un Â« Shazam nazi Â» pour lutt	http://t.co/b	http://lemde.fr/IQoZ4W
135	3078711	2013-12-05T10:0	4.0854E+17	lemondefr	1930873	241	4	indiscrets		@lemonde_	http://twitte	TRUE	TRUE	RT @lemonde_pol: #indiscrets F	http://t.co/p	http://www.lemonde.fr/les-indiscr
136	3078712	2013-12-05T10:0	4.0854E+17	lemondefr	1930873	241	16				http://twitte	FALSE	FALSE	Le PS courtise le syndicaliste Ed	http://t.co/0	http://lemde.fr/IJMsnN
137	3077853	2013-12-05T09:5	4.0853E+17	lemondefr	1930563	241	4	Redoute		@JBJacquin	http://twitte	TRUE	TRUE	RT @JBJacquin La vente de La #I	http://t.co/N	http://ecobusiness.blog.lemonde.f
138	3077854	2013-12-05T09:4	4.0853E+17	lemondefr	1930563	241	24				http://twitte	FALSE	FALSE	De l'ADN humain vieux de 400 0	http://t.co/5	http://lemde.fr/IQjDGU
139	3077855	2013-12-05T09:4	4.0853E+17	lemondefr	1930563	241	5	municipales		@manonresc	http://twitte	TRUE	TRUE	RT @manonrescan: #4 MalgrÃ© son influence croissante, le FN peine Ã constitu
140	3077856	2013-12-05T09:3	4.0853E+17	lemondefr	1930563	241	36				http://twitte	FALSE	FALSE	44 millions de personnes souffr	http://t.co/if	http://lemde.fr/IQhPh2
141	3077857	2013-12-05T09:3	4.0853E+17	lemondefr	1930563	241	11			@vincentgire	http://twitte	TRUE	TRUE	RT @vincentgiret: @lemondefr 1er mÃ©dia de presse sur le mobile en octobre a
142	3077858	2013-12-05T09:2	4.0853E+17	lemondefr	1930563	241	10			@lucbronner	http://twitte	TRUE	TRUE	RT @lucbronner: FranÃ§ois Holl	http://t.co/9	http://www.lemonde.fr/les-indiscr
143	3077859	2013-12-05T08:4	4.0852E+17	lemondefr	1930563	241	16				http://twitte	FALSE	FALSE	Roms : le maire du Var mis en c:	http://t.co/P	http://lemde.fr/IJCvXe
144	3077860	2013-12-05T07:4	4.085E+17	lemondefr	1930563	241	30				http://twitte	FALSE	FALSE	Le roi de ThaÃ¯lande appelle Ã	http://t.co/B	http://lemde.fr/IQOHYH
145	3077861	2013-12-05T07:1	4.0849E+17	lemondefr	1930563	241	15				http://twitte	FALSE	FALSE	Â« Treme Â», derniers concerts	http://t.co/6	http://lemde.fr/IJspWs
146	3077195	2013-12-05T06:0	4.0848E+17	lemondefr	1930357	241	3				http://twitte	FALSE	FALSE	L'hommage de "Fast and Furiou	http://t.co/J	http://lemde.fr/IJmfFR
147	3077196	2013-12-05T05:0	4.0846E+17	lemondefr	1930357	241	27				http://twitte	FALSE	FALSE	Centrafrique : nouveau massacr	http://t.co/B	http://lemde.fr/IPFSwL
148	3077197	2013-12-05T04:1	4.0845E+17	lemondefr	1930357	241	24				http://twitte	FALSE	FALSE	La Chine autorise une coentrepr	http://t.co/C	http://lemde.fr/IJbsvc
149	3077198	2013-12-05T03:4	4.0844E+17	lemondefr	1930357	241	8				http://twitte	FALSE	FALSE	Rob Ford aurait tentÃ© d'achet	http://t.co/t	http://lemde.fr/IPvFK7
150	3076236	2013-12-05T01:0	4.084E+17	lemondefr	1930302	241	41				http://twitte	FALSE	FALSE	Mexique : le matÃ©riel radioac	http://t.co/V	http://lemde.fr/IIP5pl
151	3076237	2013-12-05T00:2	4.0839E+17	lemondefr	1930302	241	35				http://twitte	FALSE	FALSE	Une quarantaine de baleines Ã©	http://t.co/K	http://lemde.fr/IIJIH8

```
{
    created_at: "Tue Mar 21 20:51:43 +0000 2006",
    id: 21,
    id_str: "21",
    text: "just setting up my twttr",
    source: "web",
    truncated: false,
    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: 13,
        id_str: "13",
        name: "Biz Stone",
        screen_name: "biz",
        location: "San Francisco, CA",
        url: "http://www.bizstone.com",
        description: "Co-founder of Twitter, Inc.",
        protected: false,
        followers_count: 2053993,
        friends_count: 561,
        listed_count: 18449,
        created_at: "Tue Mar 21 20:51:43 +0000 2006
        favourites_count: 914,

        utc_offset: -28800,
        time_zone: "Pacific Time (US & Canada)",
        geo_enabled: true,
        verified: true,
        statuses_count: 5117,
        lang: "en",
        contributors_enabled: false,
        is_translator: false,
        profile_background_color: "022330",
        profile_background_image_url: "http://a0.
        profile_background_image_url_https: "http
        profile_background_tile: false,
        profile_image_url:
        "http://a0.twimg.com/profile_images/34593
        profile_image_url_https:
        "https://si0.twi
        profile_banner_u              following: null,
        profile_link_col              follow_request_sent: null,
        profile_sidebar_              notifications: null
        profile_sidebar_           },
        profile_text_col           geo: null,
        profile_use_back           coordinates: null,
        default_profile:           place: null,
        default_profile_           contributors: null,
                                   retweet_count: 2157,
                                   favorite_count: 765,
                                   favorited: false,
                                   retweeted: false,
                                   lang: "en"
                                }
```

technical topics

- how deep must we go?
- other sources
- media and web capture
- search / analysis

- managing processes and data flow
- import / export / delivery
- app packaging

next steps

- improve SFM to meet diverse research, teaching, collection development needs
- meeting at GW Libraries this week
- a robust, reliable, implemented, tested, and documented application
- looking for collaborators

thanks!

gwu-libraries.github.io/social-feed-manager

@bergisjules @dchud @dankerchner @liblaura