ICPSR

SHARING DATA TO ADVANCE SCIENCE

Updates on ICPSR's Social Media Archive (SOMAR)

Libby Hemphill

ICPSR

May 2019

Talk Overview

SOMAR Research Activities

- Research and Data Practices Survey
- Reviewing Social Media Research Methods
- Implications for SOMAR

SOMAR Development

- SOMAR Goals
- SOMAR Infrastructure

Social Media Research and Data Practices Survey

- Summer 2018
- 73 responses
- 5 sections:
  1. general and demographic,
  2. data acquisition,
  3. data transformation,
  4. analysis and visualization, and
  5. data sharing and reuse

Social Media Platforms Used

Platform	%	N
Twitter	39.7%	29
Facebook	28.8%	21
Instagram	11.0%	8
Reddit	11.0%	8
Wikipedia	6.8%	5
Tumblr	5.5%	4
Other	4.1%	3
Twitch	2.7%	2
YouTube	2.7%	2
Pinterest	1.4%	1

Researcher Data Sharing Practices

Mechanism	**%**	**N**
I don't make my data available.	31.5%	23
I make my data available.	46.6%	34
In a repository or archive	15.1%	11
Through a personal website	11.0%	8
Through journal or conference site	8.2%	6
Through a University affiliated website	6.8%	5
Through a third party data provider	5.4%	4
Other	15.1%	11

Methods Sections

Reviewed 40 articles in 4 high-impact journals:

1. No one provided access to data
2. No one provided enough detail to replicate data collection
3. Data collected computationally (e.g., Python scripts) or through third-parties (e.g., Crimson Hexagon)

How is social media like other social science data?

1. Researchers worry about getting scooped
2. Preparing data for reuse takes a lot of effort
3. Found data requires special manipulation and documentation

What makes social media data special?

1. Data properties: structure, scale, speed
2. Data practices: finding, curating, sharing, and storing
3. Ethics: private owners, PII

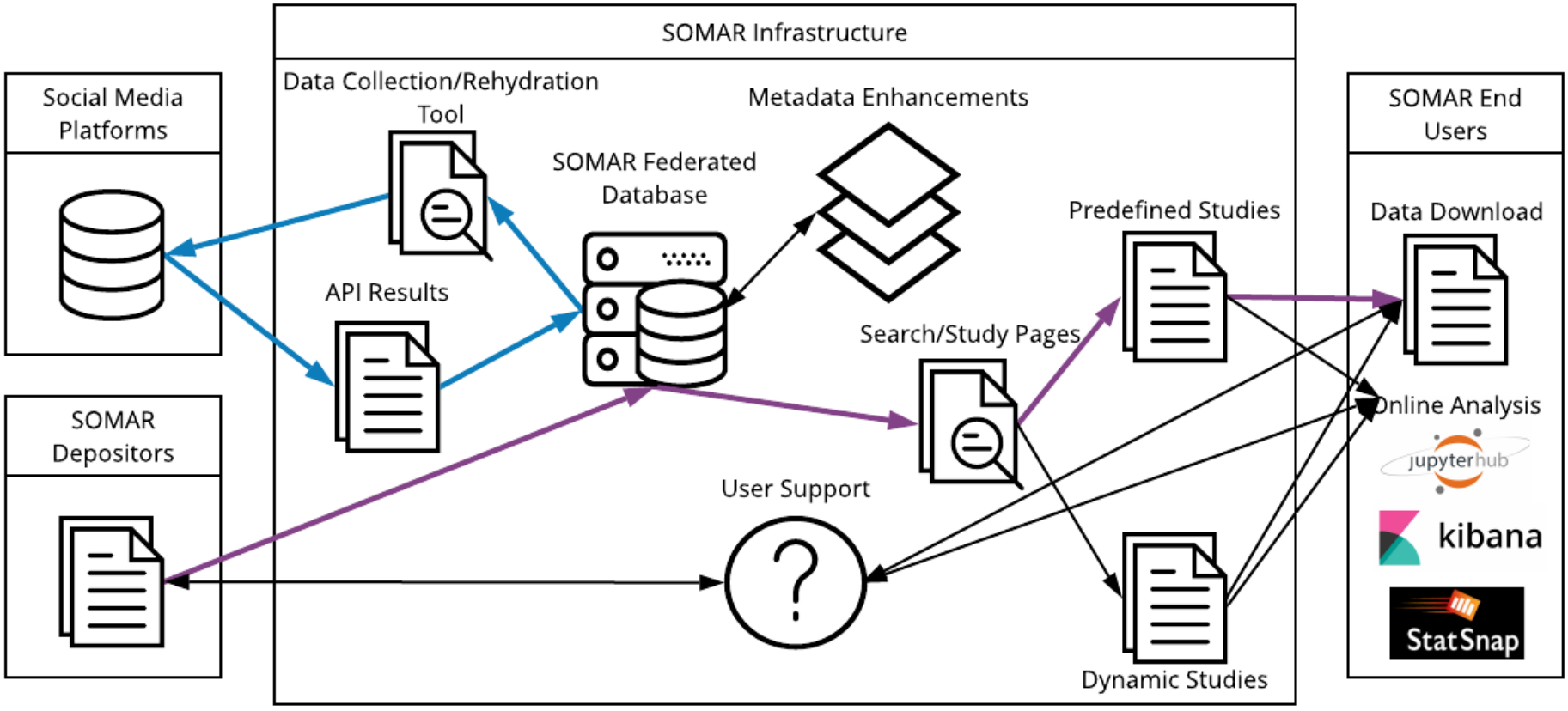
Implications for SOMAR

1. richer documentation of provenance,
2. explicit documentation of the terms and conditions for acquiring the data, and
3. software and/or scripts used to acquire and manipulate the data
4. what constitutes a dataset
5. observation-level metadata enhancements and declarations

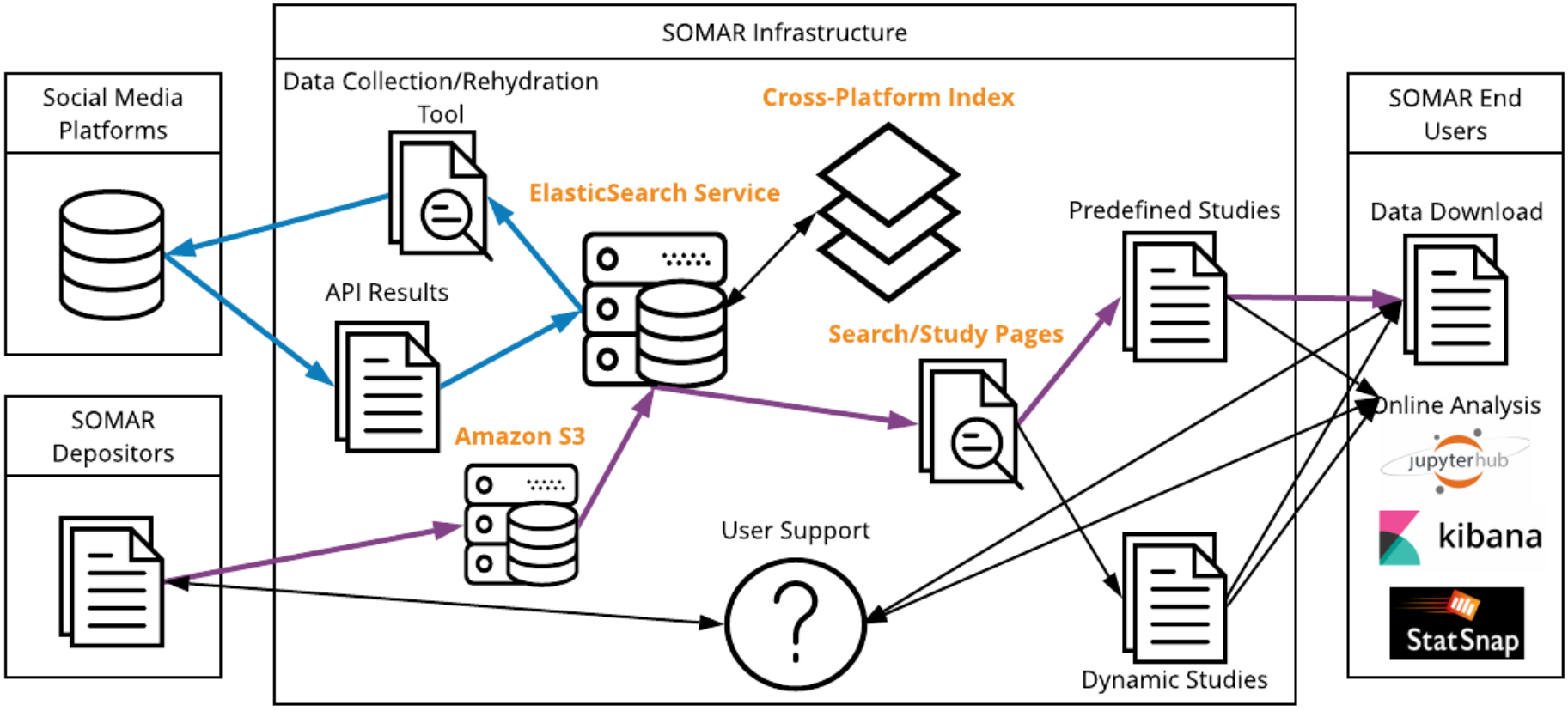
SOMAR Goals

1. Accept data from multiple platforms
2. Archive both data and code
3. Balance data terms, restrictions, and researcher needs
4. Scale responsibly

SOMAR Technical Infrastructure



SOMAR Technical Infrastructure



Ethical Dilemmas

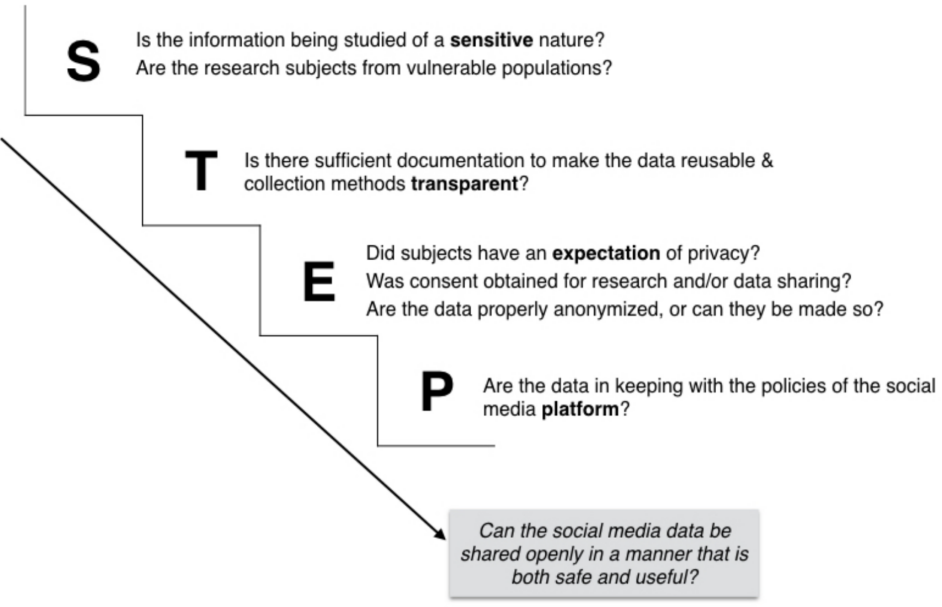


Figure 1. Visualization of the STEP framework for curating social media data.

Mannheimer and Hull 2018

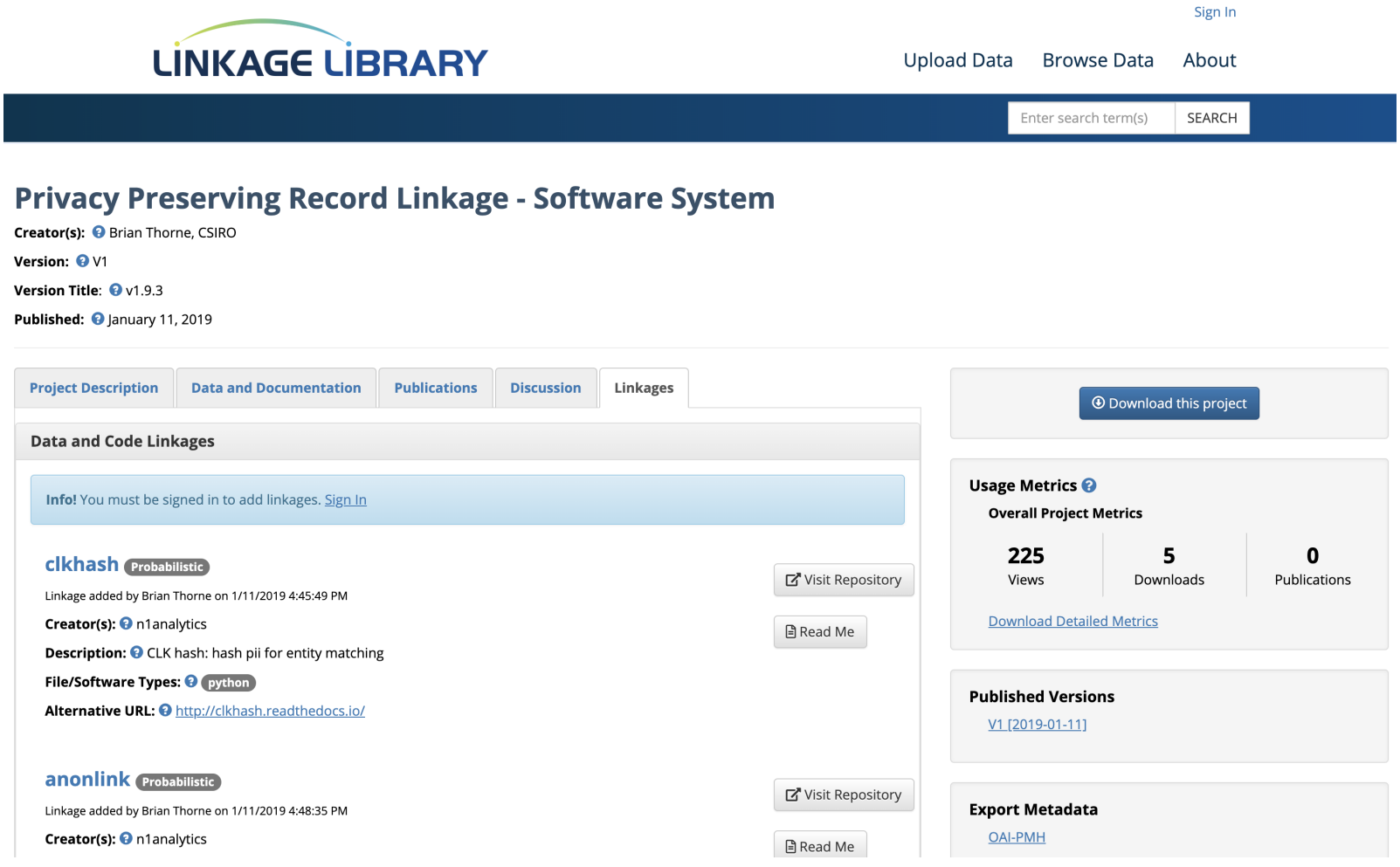
Table 4. "How Would You Feel If a Tweet of Yours Was Used in a Research Study and . . ." ($n = 268$).

	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
. . . you were not informed at all?	35.1%	31.7%	16.4%	13.4%	3.4%
. . . you were informed about the use after the fact?	21.3%	29.1%	20.5%	22.0%	7.1%
. . . it was analyzed along with millions of other tweets?	2.6%	18.7%	25.5%	30.0%	23.2%
. . . it was analyzed along with only a few dozen tweets?	16.5%	30.3%	24.0%	20.2%	9.0%
. . . it was from your "protected" account?	54.9%	20.5%	13.8%	6.0%	4.9%
. . . it was a public tweet you had later deleted?	31.3%	32.5%	20.5%	10.4%	5.2%
. . . no human researchers read it, but it was analyzed by a computer program?	2.6%	14.3%	30.5%	32.3%	20.3%
. . . the human researchers read your tweet to analyze it?	9.7%	27.6%	25.0%	25.4%	12.3%
. . . the researchers also analyzed your public profile information, such as location and username?	32.2%	23.2%	21.0%	13.9%	9.7%
. . . the researchers did not have any of your additional profile information?	4.9%	15.4%	25.1%	34.1%	20.6%
. . . your tweet was quoted in a published research paper, attributed to your Twitter handle?	34.3%	21.6%	21.6%	13.1%	9.3%
. . . your tweet was quoted in a published research paper, attributed anonymously?	9.0%	16.8%	26.5%	28.4%	19.4%

Note. The shading was used to provide a visual cue about higher percentages.

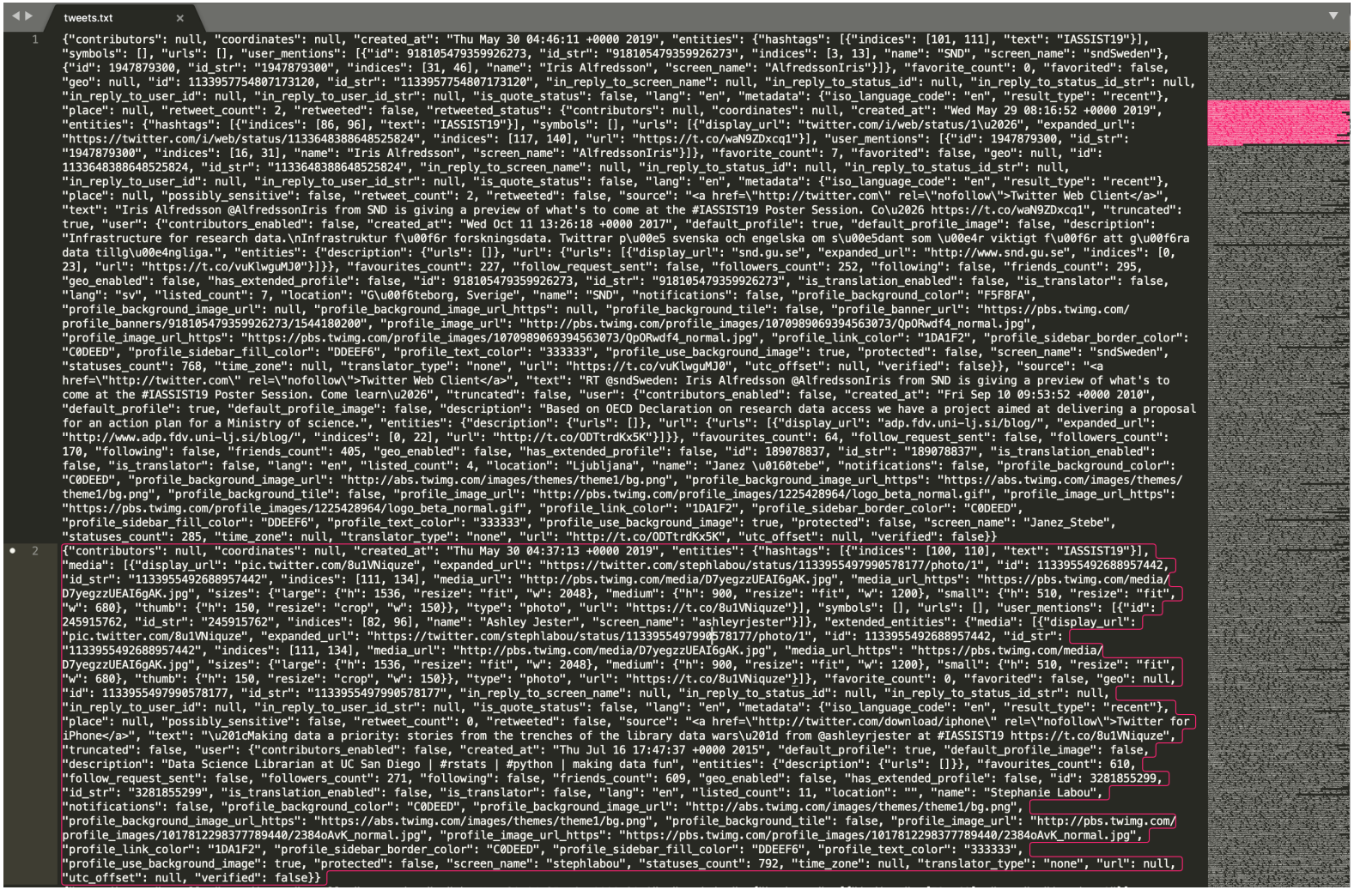
Fiesler and Proferes 2018

Archiving Data + Code



Archiving Data + Code

What does data look like?



Pretty printed

Archiving Data + Code

What does code look like?

- twint - command line
- open humans - Jupyter notebook
- hydrator - desktop app option
- purpletag - command line

Metadata

OPEN ICPSR Find Data Share Data openICPSR Repositories GO libbyh@umich.edu

Find Data / Who Did US Congress Retweet in 2017

Who Did US Congress Retweet in 2017

Principal Investigator(s): Libby Hemphill, University of Michigan; Angela Schopke, University of Michigan; Caroline Hodge, University of Michigan; Chris Bredernitz, University of Michigan

Version: V1

Version Title: Initial deposit

Name	File Type	Size	Last Modified
retweets_2017_complete.csv	text/csv	427.7 MB	01/28/2019 06:27:AM

DOWNLOAD THIS PROJECT

Project Citation:

Hemphill, Libby, Schopke, Angela, Hodge, Caroline, and Bredernitz, Chris. Who Did US Congress Retweet in 2017. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-01-28. https://doi.org/10.3886/E108303V1

Persistent URL: http://doi.org/10.3886/E108303V1

Project Description

Summary: This dataset includes the retweets posted on Twitter by accounts associated with members of the US Congress in 2017. The list of accounts combines two sources:

- Justin Littman's list (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UIVHQR)
- The United States project list (https://github.com/unitedstates/congress-legislators)

Tweets were collected using Twitter's Search API through the twitter_user_collector Python script (https://github.com/casmlab/twitter_user_collector).

Scope of Project

Subject Terms: Congress; twitter

Geographic Coverage: United States

Usage Metrics

Overall Project Metrics

217	14	0
Views	Downloads	Publications

Download Detailed Metrics

Published Versions

V1 [2019-01-28]

Export Metadata

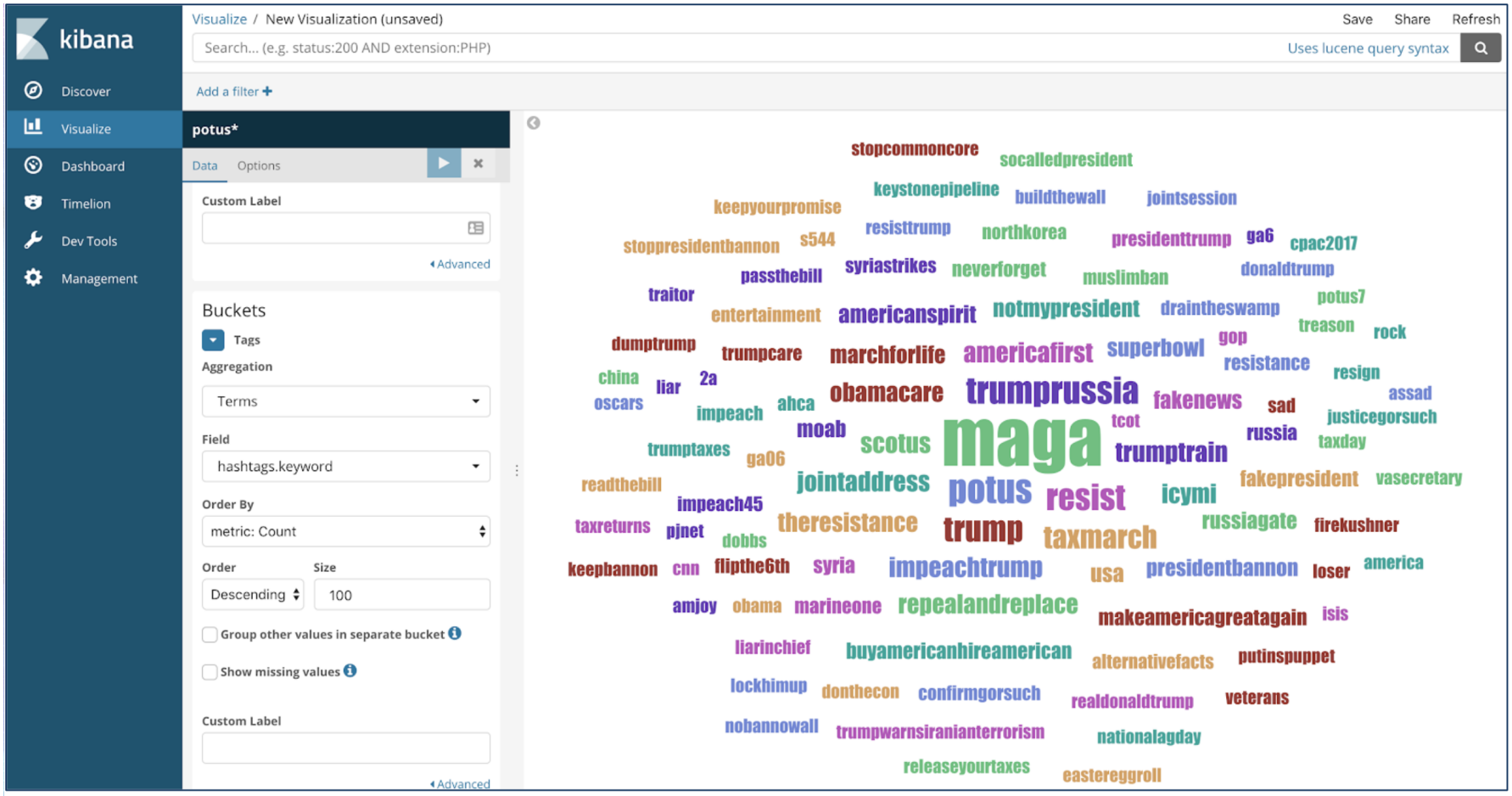
Dublin Core

DDI 2.5

DDI 3.1

Congress's retweets example

Integrating Search + Analysis



SOMAR Questions

1. How should SOMAR prioritize its goals? (e.g., replication, preservation, active collection)
2. How would users most likely search for data in SOMAR?
3. What tools exist already on your campus for facilitating social media data access? (e.g., Crimson Hexagon, Sysomos, Python/R consulting)
4. Which ethical considerations should guide our decisions, and how do they lead to different technical outcomes?