Developing a Social Media Archive at ICPSR

Libby Hemphill
University of Michigan
Ann Arbor, MI
libbyh@umich.edu

Susan H. Leonard
University of Michigan
Ann Arbor, MI
hautanie@umich.edu

Margaret Hedstrom
University of Michigan
Ann Arbor, MI
hedstrom@umich.edu

ABSTRACT

Social media are implicated in many of contemporary society's most pressing issues, from influencing public opinion, to organizing social movements, and identifying economic trends. Increasing the capacity of researchers to understand the dynamics of such social, behavioral and economic phenomena will depend on reliable, curated, discoverable and accessible social media data. To that end, ICPSR will develop a new archive of curated datasets, workflows, and code for use by social science researchers for the empirical analysis of social media platforms, content, and user behavior. The goal is to provide a user-friendly, large-scale, next-generation data resource for researchers conducting data-intensive research using data from social media platforms such as Facebook, Twitter, Reddit, and Instagram. In our presentation, we will explain SOMAR's goals and structure and discuss opportunities for collaboration.

CCS CONCEPTS

• **Information systems → Data management systems**;

KEYWORDS

social media archiving, collection building, community building

ACM Reference Format:
Libby Hemphill, Susan H. Leonard, and Margaret Hedstrom. 2018. Developing a Social Media Archive at ICPSR. In *Proceedings of Web Archiving and Digital Libraries (WADL'18).* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

An archive for social media data will enable researchers to discover reusable social media datasets, provide a means for evaluating and/or replicating research based on social media data, and enable new insights, longitudinal studies, or comparative analyses that are nearly impossible today. Common, transparent, and reproducible approaches to privacy protection, linkage methodology, and analytical tools for these data will help ensure that research using social media data meets the highest scientific and ethical standards, and therefore gains the legitimacy necessary to advance the underlying science to its full potential.

The Social Media Archive (SOMAR) will bring together social media datasets as a corpus with associated services and resources to aid researchers in further interacting with and mining the data. This will enable extension of original findings and the creation of new knowledge, leading to a greater return on the original investment in the data. Research has shown strong and consistent evidence that data sharing, both formal and informal, increases research productivity across a wide range of publication metrics and that formal data archiving, in particular, yields the greatest returns on investment with an increased number of publications resulting when data are archived [3, 4].

We currently focus our efforts on addressing four communities of researchers: those who (1) study social media use specifically, (2) leverage social media data to understand people and society more generally, (3) study social science methods, and (4) investigate new methods for curation, publication, confidentiality and quality assessment, and long-term management of research data. One of the primary benefits of the archive is that it enables historical and longitudinal analysis. In the absence of an archive, these questions have been explored in specific, isolated, historical, social, political, and technical moments, and SOMAR's federation and long-term availability of data enables research across and between those moments.

2 TECHNICAL OVERVIEW

SOMAR will be designed to house a variety of data products related to social media research, supported by a newly developed and sustainable infrastructure. We anticipate that there will be some social media data analysis projects for which the data themselves cannot be deposited – for example, cases in which platform terms of service prohibit data sharing. As these are often exactly the instances where transparency or replicability are lacking, ICPSR will in these cases archive data workflows and code that enable users to replicate the data collection, transformation, and analysis procedures researchers' followed.

Federating data through a shared archive will result in more opportunities for comparative and historical analyses, higher quality user experience, less duplication of effort, and lower overall costs. By capturing metadata such as included hashtags and dates, SOMAR also makes it possible to generate new datasets by searching across deposits. For instance, if the same hashtag appears in multiple Twitter datasets, users could generate a dataset of those tweets even if the hashtag wasn't an original search term in any of the datasets.

Figure 1 provides an overview of the SOMAR system. There are many possible paths through the system, but the most common is likely that a user deposits data (and associated files) such as "dehydrated" data or unique identifiers of social media content (e.g., tweet ids). Often, platform terms of service dictate what data users are allowed to deposit. Twitter, for instance, allows users to share the IDs of tweets but not the tweets themselves. Common practice among researchers is then to share the list of tweet IDs included in

WADL'18, June 2018, Fort Worth, Texas USA

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

a study and some link to code that would re-collect those tweets through the Twitter API [1, 5]. For instance, the Beyond the Hashtags dataset [2] contains roughly 40 million tweet ids; the website where it resides provides Python code for rehydrating the data through the Twitter API. If the data a user deposits is dehydrated, then the SOMAR system rehydrates that data by querying the platform's API (see the blue loop in Figure 1) and stores the complete data on its servers.

The curation team then uses both the data deposited and rehydrated data to create metadata enhancements (e.g., provenance, description of the platform at the time of collection, dates). Some enhancements such as expanding shortened URLs and using consistent case for hashtags and mentions [1] are straightforward and can be accomplished programmatically while others, such as disclosure risk review, require human labor.

SOMAR end users can then access data through pre-defined studies where the data they download is the same (plus metadata enhancements) as the data deposited. For instance, in the Beyond the Hashtags example, users would be able to download the list of tweet IDs or to interact with them in JupyterHub. By federating rehydrated datasets, SOMAR also enables end users to create dynamic studies by querying the entire SOMAR database and retrieving results that include data from multiple studies. For instance, they may query for all data with a certain date stamp or containing a particular set of terms and receive subsets of Researcher A's and Researcher B's studies. These dynamic studies may also include data from multiple platforms (e.g, Twitter and Reddit). End users may interact with the data through download or through JupyterHub. In this overview "data" refers to all data, documentation, code, etc.

ICPSR provides user support across the system, but most user contact occurs around deposit, download, and JupyterHub.

3 GOVERNANCE

A Steering Committee co-led by Libby Hemphill and Margaret Levenstein will set the direction of SOMAR and have final say on features and design. This steering committee will provide a mechanism for communication and governance to ensure that the needs and perspectives of the different disciplines involved in SOMAR are fully considered. We have recruited researchers from each of the scientific communities described above who regularly grapple with the challenges associated with social media data, such as managing the scale and velocity of social media data, understanding platform terms of service, and handling personally identifiable information.

4 CONCLUSION

As we begin developing each of SOMAR's components, ICPSR is interested in feedback from and collaboration with other researchers collecting, managing, and using social media data. Current opportunities include participating in our study of social media data management practices, depositing social media datasets to seed SOMAR and inform its development, and researching ways to link social media data with other data types (e.g., census, surveys) without compromising individual users.

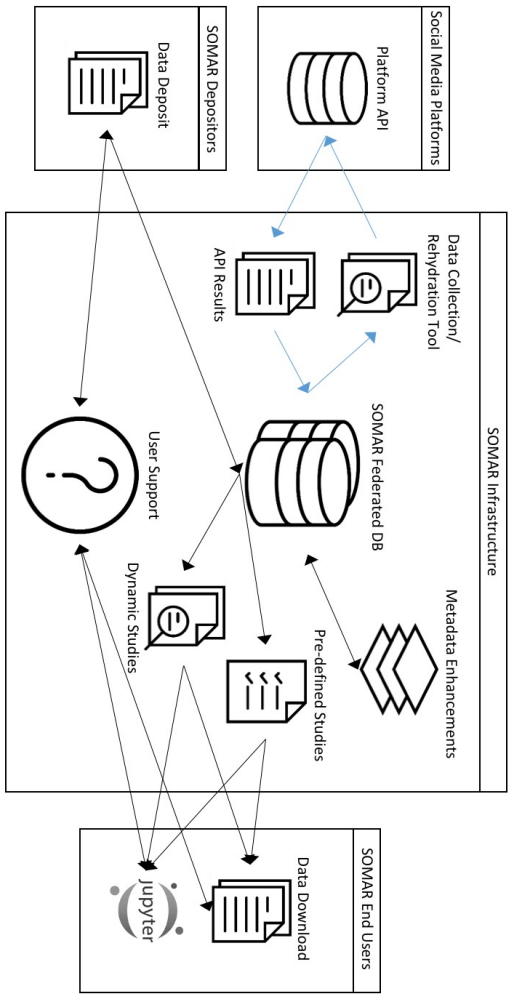

Figure 1: SOMAR System Overview.

REFERENCES

[1] Kevin Driscoll and Shawn Walker. 2014. Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *Int. J. Commun. Syst.* 8 (2014), 20.
[2] D Freelon. 2017. Beyond the Hashtags Twitter data. (Jan. 2017).
[3] Amy M Pienta, George C Alter, and Jared A Lyle. 2010. The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. (Nov. 2010).
[4] Heather A Piwowar, Roger S Day, and Douglas B Fridsma. 2007. Sharing detailed research data is associated with increased citation rate. *PLoS One* 2, 3 (March 2007), e308.
[5] Katrin Weller and Katharina E Kinder-Kurlanda. 2015. Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research. In *Standards and practices in large-scale social media research. Oxford: International Conference on Web and Social Media.*