Hacking the Storage and Preservation of Social Media Data

Anatoliy Gruzd[1], Jenna Jacobson[2], Elizabeth Dubois[3]
[1]Ryerson University
[2]University of Toronto
[3]University of Oxford

Abstract

The growing availability of social media has afforded researchers the ability to conduct large-scale research projects using social media data. Social media platforms will come and go, but what is constant is the fact that in order to foster the sharing of data and encourage research innovation, there is a pressing need for the research community to develop a strong set of data stewardship principles, standards, and protocols around social media data preservation. The hackathon will bring overlapping research communities of Information scholars together to identify the major challenges, opportunities, and possible interventions to address the preservation and storage of social media data.
Keywords: social media; data curation; data stewardship; big data
doi: 10.9776/16551

Contact: gruzd@ryerson.ca, jenna.jacobson@mail.utoronto.ca, dubois.elizabeth@gmail.com

1 Introduction

The increasing availability of social media data is bringing large and dynamic datasets to fields, such as social sciences, that have not been traditionally considered data-driven. This has allowed scholars in those fields to study human condition on an unprecedented scale, but a culture of stewardship is desperately needed for social media data.

Social media platforms will come and go, but what is constant and important is the fact that in order to foster the sharing of data and encourage research innovation, there is a pressing need for the research community to develop a strong set of data stewardship principles, standards, and protocols around social media data preservation.

Research data stewardship studies have primarily focused on more “technical” and “procedural” issues related to handling scientific data: what are the best ways to preserve, retrieve, publish and share research data and metadata with collaborators, peers and the public? Considering that social media data are inherently “social,” and thus inevitably contain sensitive information about social media users and their social networks, this hackathon will move beyond technical issues and will also identify and classify various social and ethical factors of preserving social media data as research data.

This is the first academic workshop part of a larger research project on *social media data stewardship*: the collection, storage, use, reuse, analysis, publishing and preservation of social media data. The project is the first known, major undertaking to apply the emerging notion of data stewardship to social media data - data contributed by social media users directly (such as posts, photos, videos) or automatically recorded by online services and devices that users use (for example, information about who talks to whom or users’ location as reported by their mobile devices).

2 Social Media Data Preservation Hackathon

In this hackathon, we address the growing need for storage, preservation, and reuse of social media data. In academic research, there is often a need to validate other researchers’ scholarly publications and a desire to reuse datasets for other research; however, publishers and libraries are currently ill-equipped to deal with the vast and growing amount social media data used in academic research. There have been some attempts to create online repositories where users can share datasets (e.g. FigShare, ICWSM), but the data is often very limited or outdated, or there is a need to a sign a limited user agreement.

We ask teams of Information researchers to imagine: what would an open data repository of social media data look like? Can we build a social media data library? Does the institutional repository model apply to social media data?

Teams can address a single challenge, a collection of challenges, or consider a new challenge. For example:

Data Presentation and Sharing – e.g. What are the main issues associated with data sharing?
Data Storage and Preservation – e.g. What are the main strategies to ensure data reuse?

Policy Barriers – e.g. How can organizations negotiate with social media platforms to get the rights to preserve/share social media data?*

Ethical/Privacy Considerations – e.g. How do we ensure the privacy of personal data?

Technical Restrictions – What metadata schemas and what formats should be used?

Scalability Issues – e.g. How do we address scalability issues due to the immense size of some social media datasets?

*As a resource for the participating teams, we have put together an interactive website that highlights the restrictions of access, storage, and sharing on the four most popular social media platforms: Facebook, Twitter, Instagram, and Reddit. The resource is available at: http://api.socialmediadata.org

3 Workshop Activities

The hackathon will bring overlapping research communities of Information scholars together to share ideas in identifying the major challenges, opportunities, and possible future initiatives to address the preservation of social media data in academic publications.

3.1 Length of the Event

"Hacking the Storage and Preservation of Social Media Data" will be a half-day event that will generate competitive and collective action, coupled with deliberate reflection.

3.2 Agenda

Introduction to Social Media Data Stewardship – 45 minutes

- Gruzd, Jacobson, and Dubois will present on the challenges of social media data stewardship with a specific focus on data storage and preservation
- Participants' introduction and description of area of research

Break – 15 minutes

Hackathon – 90 minutes

- Form teams of approximately 5-8 people
    - Individual participants will gather to form teams with complementary skills sets
    - Participants will also be encouraged to join the hackathon as a pre-formed team

Break – 15 minutes

Team Presentations – 45 minutes

Collective Discussion & Closing – 30 minutes

- Participants will be encouraged to collectively reflect on the presentations
- Discussion of next steps, challenges, and the development of a research network

4 Significance to the Field

4.1 Purpose and Intended Audience

Information researchers are uniquely situated to address the issues of social media data stewardship by bringing in expertise from various Information subfields including: information science, data curation, library science, informatics, etc. The overarching goal of the session is to develop a specialized network of international Information scholars who are interested in collaboratively working to address the challenges of large-scale social media preservation.

4.2 Dissemination

The website *socialmediadata.org* will be developed as a place where the organizers will disseminate the results of the event as a follow-up. The website will also serve as an online place where researchers from around the world can come together after the conference to continue to exchange ideas and build

research collaborations. Throughout the session, participants will be encouraged to use the official conference hashtag *#iconf16* and the session-specific hashtag *#smdataorg*