Selected Papers of #AoIR2019:
The 20th Annual Conference of the
Association of Internet Researchers
Brisbane, Australia / 2-5 October 2019

THE QUT DIGITAL OBSERVATORY PROJECT: BUILDING A TRUSTED DATA INFRASTRUCTURE FOR SOCIAL MEDIA RESEARCH

Marissa Takahashi
Sam Hames
Elizabeth Alpert
Digital Observatory, Institute for Future Environments, Queensland University of Technology, Brisbane, Australia

Axel Bruns
Digital Media Research Centre, Queensland University of Technology, Brisbane, Australia

Introduction

Trust is fragile. The 2018 Facebook and Cambridge Analytica debacles highlighted how data harvested from social media platforms can be used not only for commercial purposes but also for political manipulation. This incident and the widespread discussion around it further demonstrated the following issues: unethical data collection enabled by a platform; unethical use of data for corporate and political interest; and unethical data sharing by an academic.

Research needs to be credible to maintain social license. Data is the lifeblood of research. For research to remain credible, research needs to remain fundamentally ethical and research methods comprising data collection and data analysis need to be robust, transparent, repeatable, and auditable. Such methods alone cannot create credibility, but research data infrastructure design and implementation can provide a foundation for credibility by addressing these fundamental processes.

Social science research has traditionally relied on data collection methods such as surveys, interviews, and ethnographic observations. However, an increasing proportion of human life is being mediated by online platforms, with 2.3 billion active users on Facebook and 326 million active users on Twitter (Statista 2019). Social media data collection and analysis have become imperative for researchers interested in various phenomena playing out in these new media. This paper discusses the current state and

Suggested Citation (APA): Takahashi, M., Hames, S. Alpert, E., & Bruns, A. (2019, October 2-5). *The QUT Digital Observatory Project: Building a Trusted Data Infrastructure ffor Social Media Research.* Paper presented at AoIR 2019: The 20th Annual Conference of the Association of Internet Researchers. Brisbane, Australia: AoIR. Retrieved from http://spir.aoir.org.

issues of social media data collection and describes the Digital Observatory's approach to establishing a credible and trusted research data infrastructure.

Social Media Data Collection and Methods

Researchers have employed various means for collecting social media data. This ranges from obtaining data from platforms like Facebook and Twitter via an application programming interface (API) or third-party tools to collecting data manually (e.g. copy-and-paste into spreadsheets or other databases), to create a corpus for their research. Alongside data collection, there are also a variety of emerging analytical methods developing in conjunction with traditional approaches.

There are many research studies that use Twitter data. Approaches in these studies can be classified as: platform studies, hashtag studies, and population studies. Platform studies look at the Twitter platform as the object of study and examine its use, adoption, and affordances. Hashtag studies use datasets collected by specifying the relevant hashtags and are easy to collect. However, hashtag datasets are unable to capture the wider communicative activities around the topic or event described by the hashtag. Bruns (2018) describes these studies as "analogous to listening to only one side of a multi-sided phone conversation."

Population studies address the limitations of hashtag studies by collecting data from a curated population, for example within a specific geographic location. Studies using collections such as the Australian Twittersphere (Bruns, et al. 2014) provide the necessary context and benchmark information that is not available from hashtag studies. A major challenge with such large-scale population studies is the massive data collection effort needed.

Social Media Data Collection Issues

The credibility and quality of research output depend on the quality of data collected and the suitability of the analytical methods. There are policy and regulation issues relating to data collection from digital platforms. First, many platform providers operate as closed domains. Access to these platforms' APIs has become increasingly difficult, especially after the Cambridge Analytica debacle. Second, governments have established stricter regulations such as the General Data Protection Regulation (GDPR) in the European Union in 2018 to protect citizens' privacy and rights over their personal data. There are also existing laws on reporting breaches and data sovereignty.

Furthermore, there are methodological issues relating to social media data collection. First, the fractured approach to data collection, typically building on one-off processes using ad-hoc software tools, is neither robust nor repeatable. The software tools effectively function as black boxes, and researchers are therefore not privy to their limitations (e.g., legal, technical, and terms of service) and how these affect the data quality. Second, when the collection method is not transparent, the characteristics of the data collected - in terms of randomness, representativeness, and completeness - cannot be determined. These unknowns will have a major influence on the analytical methods that can be applied and the quality of insights that can be derived. Lastly, while

there is the option of purchasing Twitter data via GNIP, the costs are usually not affordable for academic researchers (Gaffney & Puschmann 2014).

To address the data collection issues mentioned above, the TrISMA (Tracking Infrastructure for Social Media Analysis) project designed, built, and operated a 'big social data' infrastructure for multiple platforms (Bruns 2018). There were four challenges with the TrISMA project: technical, coordination, maintenance, and methods. The technical challenges involved in gathering, processing, and storing such large datasets were significant. In addition, the multi-institutional nature of the project naturally generated substantial coordination overheads. Additional challenges include the maintainability of the complex architecture necessary to operate such a facility, the lack of documentation common to prototypes, and changing platform APIs. Finally, the analytics methods required to use the TrISMA data remain emergent and experimental.

Building a Trusted Data Research Infrastructure

The Digital Observatory (DO), building on the TrISMA project, aims to address the challenge of building a trusted platform that makes it possible to collect data responsibly, ethically, and securely. The DO aims to accomplish this via a two-pronged approach: technological and procedural. The technological approach means making conscious design decisions to build data collectors that are reliable, maintainable, and respectful of the platforms they gather from, and of the authors of data on those platforms. The procedural component involves structuring the DO processes around our guiding principles: being an active facilitator and not just a passive data provider.

The DO takes an infrastructure approach to address the issues arising from the fragmented approach to research methods mentioned before, and to ensure robustness, transparency, repeatability, and auditability. Firstly, it has learned from TrISMA and redesigned and rebuilt that prototype into production-grade software using professional software development best practice, increasing reliability and maintainability. Secondly, the centralised approach to data collection provides scalability and reduces the duplication of data collection and processing work that happens across research institutions. Finally, the provision of continuous support for the platform, which is necessary but tangential to the conceptual workflow of researchers, allows researchers to focus more on the intellectual challenges of their research.

This paper will outline the successes as well as the challenges arising in the design and development of the new Digital Observatory platform. We also outline the risks and strategies for mitigating these risks. We anticipate that once the DO platform has been developed and is in a steady state of operation, the next major challenge will be to extend the platform with innovative analytical methods.

References

Bruns, A., Burgess, J., & Highfield, T. (2014). A 'big data' approach to mapping the Australian Twittersphere. In *Advancing digital humanities* (pp. 113-129). Palgrave Macmillan, London.

Bruns, A. (2018). A multi-institutional approach to 'big social data': The TrISMA project. Paper presented at AoIR 2018, Montréal, Canada.

Gaffney, D., & Puschmann, C. (2014). Data collection on Twitter. In Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C., eds. *Twitter and society*. 55-67. Peter Lang, New York.

Statista. (2019). Most popular social networks worldwide as of January 2019, ranked by number of active users (in millions). Retrieved 28 February 2019, from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/