Social Media Data in Research: Provenance Challenges*

David Corsar, Milan Markovic, and Peter Edwards

Computing Science, University of Aberdeen, Aberdeen, UK
{dcorsar,milan.markovic,p.edwards}@abdn.ac.uk

Abstract. In this paper we argue that understanding the provenance of social media datasets and their analysis is critical to addressing challenges faced by the social science research community in terms of the reliability and reproducibility of research utilising such data. Based on analysis of existing projects that use social media data, we present a number of research questions for the provenance community, which if addressed would help increase the transparency of the research process, aid reproducibility, and facilitate data reuse in the social sciences.

Keywords: Provenance, social media, research process

1 Introduction

The social science research community faces challenges associated with the reliability, statistical validity, and generalizability of data obtained from social media [Tuf13], which may raise questions about the validity of research based on such data and hinder data reuse [fECoD13]. Provenance has previously been used to support audit, verification, and reproducibility in a number of domains [Mor11,CFLV12]; as such, we argue that documenting the provenance of social media data and its subsequent analysis could help address the challenges faced by the social sciences - by increasing the transparency of the research process, and supporting assessment of the analytical methods used.

2 Case Study - *Tweeting Transport*

To investigate this application of provenance we have analysed a number of projects that utilised social media data; one of these will now be described in order to provide context for the research questions in Section 3. The *Tweeting Transport* project [CYG+15] explored how Twitter[1] is used to provide transport information during major events, focusing on the 2014 Commonwealth Games[2].

* The work described here was funded by a grant from the United Kingdom's Economic and Social Research Council Social Media - Developing Understanding, Infrastructure & Engagement (ES/M001628/1).

[1] http://www.twitter.com

[2] http://www.glasgow2014.com

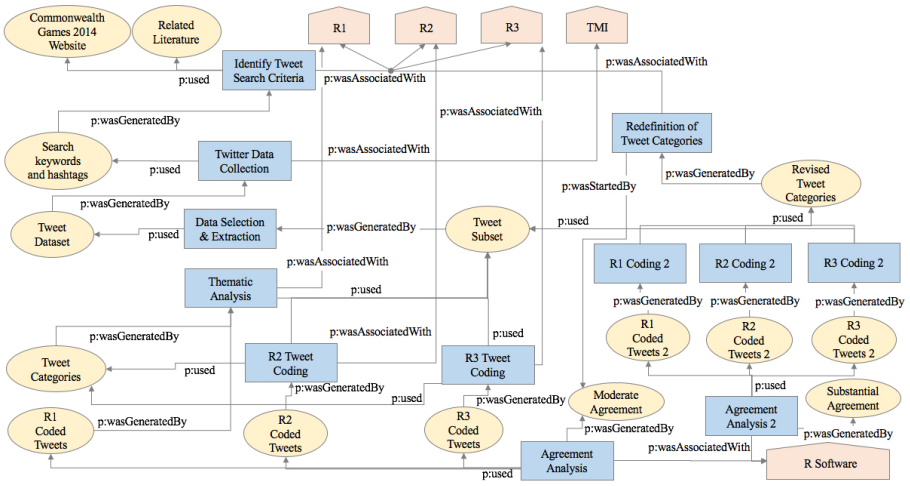

Fig. 1. PROV representation of the *Tweeting Transport* project.

Fig. 1 provides a PROV [MGC+15] representation of the *Tweeting Transport* project. A dataset of tweets relating to transport disruption during the 2014 Games was created using TMI[3], a tool developed to monitor Twitter, and to store and export Tweets to CSV files for analysis. TMI was configured to capture tweets containing at least one of 331 keywords or hashtags, as well as tweets authored by eight different user accounts. These criteria were based on a review of travel information published via the official Games website[4] and a review of the wider transport disruption literature. Data were collected one week before the event, during the Games, and for one week afterwards (July/August 2014).

Three types of analysis were subsequently performed to understand the kinds of travel information provided, and the Twitter users who disseminated this content. Here we summarise the first of these, which focused on Retweets and replies in response to Tweets sent by the official travel information Twitter account, @GamesTravel2014. The analysis involved thematic coding of each tweet by one researcher (R1 in Fig. 1), which categorised each tweet based on its content. These categories were used by two additional researchers (R2 and R3 in Fig. 1) to code the same data, which resulted in moderate agreement between the coders (as computed by the Fleiss Kappa implementation of the R tool[5]). Following discussions between the researchers, the categories were redefined, and the dataset recoded, resulting in substantial agreement. Seven types of travel information that were shared via Twitter were identified.

[3] https://github.com/SocialJourneys/TMI
[4] http://www.glasgow2014.com/your-games/travel-and-transport
[5] https://www.r-project.org/

3 Research Questions

The approach to data collection and analysis described above is typical of such projects; [BT14] presents a taxonomy of social media providers, data types, and access mechanisms; data cleaning, tagging, and storing activities; and techniques and tools commonly used with such data. Documenting these various aspects of social media analytics forms the basis of our research questions.

RQ1 - What characteristics of social media data should be captured to facilitate transparency, and reproducibility of such research?

We argue that it is necessary to capture aspects of *why*, *how*, *where* and *when* [CCT09] data provenance. *Why* characteristics capture both why the dataset was created, and why each datum was included; *how* characteristics define how the data were acquired, for example, via the Twitter Stream API[6] and/or tools such as TMI; *where* characteristics define the source of the data, for example Facebook[7] or a third party service such as the Gnip[8] enterprise platform; and *when* defines both the temporal coverage of the data, and when collection took place. These are necessary to allow others to understand the data (including restrictions on reuse due to license conditions), and to understand how to reproduce the dataset if necessary.

RQ2 - How can existing provenance models be employed to record analysis of social media data?

The analysis (and associated stages, such as data preparation) can be viewed as a set of activities that use, generate, and exchange information. The *Tweeting Transport* project also illustrates why it will be necessary to capture the different agents that were involved in these activities (as three researchers conducted the Tweet coding activity independently). This is consistent with the *process flow* view of provenance [MGC+15], which PROV is capable of capturing. While models, such as PROV-SAID[9] extend PROV with the ability to capture information diffusion within social media platforms, further extensions are required to capture different types of analysis, such as thematic coding and recursive abstraction.

RQ3 - What information should the provenance record contain to facilitate transparency and reproducibility of research that utilises social media data?

This question considers the appropriate level(s) of granularity required. For example, is it necessary for the provenance record of the *Tweeting Transport* project to contain all of the information regarding the revision of the initial Tweet categories (as in Fig. 1), or does a description of the revised categories and coded Tweets provide sufficient detail to allow others to reproduce the research?

RQ4 - How can the provenance of social media analysis be captured?

One obvious approach here would be construction of a software tool, able to guide a researcher through creation of a description of their data and analytical processes. However, previous experience in the *ourSpaces* Virtual Research

[6] `https://dev.twitter.com/streaming/overview`
[7] `https://www.facebook.com`
[8] `https://gnip.com/`
[9] `http://semweb.mmlab.be/ns/prov-said/`

Environment [EPE+12] indicates that the descriptions obtained in this way are likely to be limited, as few users will provide details beyond the minimum required when describing, for example, a dataset. As such, we argue that it will be necessary to develop automated solutions that attempt to infer or reconstruct (parts of) the provenance record by, for example, examining data files generated by popular qualitative data analysis tools such as NVivo[10].

4 Future Work

As part of our investigation of these research questions, we are currently developing the model extensions necessary to enable capture of the provenance of research that uses social media data. Following this, we plan to develop a software tool that supports creation of provenance expressed using the new model; the tool will be evaluated by application to our case study projects. We are also developing a set of guidelines that will support research data archives to obtain the appropriate information from those conducting research using social media data, to provide others with greater understanding of the research undertaken, knowledge of how to verify, repeat and/or reproduce the research, and to facilitate greater data reuse.

References

[BT14] B. Batrinca and P. C. Treleaven. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1):89–116, 2014.

[CCT09] J. Cheney, L. Chiticariu, and W.-C. Tan. Provenance in databases: Why, how, and where. *Found. Trends databases*, 1(4):379–474, April 2009.

[CFLV12] J. Cheney, A. Finkelstein, B. Ludascher, and S. Vansummeren. Principles of Provenance. *Dagstuhl Reports*, 2(2):84–113, 2012.

[CYG+15] C. Cottrill, G. Yeboah, P. Gault, J. D. Nelson, J. Anable, and T. Budd. Tweeting transport: Examining the use of twitter in transport events. In *Proceedings of the 47th Annual UTSG Conference*, 2015.

[EPE+12] P. Edwards, E. Pignotti, A. Eckhardt, K. Ponnamperuma, C. Mellish, and T. Bouttaz. ourSpaces–Design and Deployment of a Semantic Virtual Research Environment. In *The Semantic Web–ISWC 2012*, pages 50–65. Springer, 2012.

[fECoD13] Organisation for Economic Co-operation and Development. New data for understanding the human condition. Technical report, February 2013.

[MGC+15] L. Moreau, P. Groth, J. Cheney, T. Lebo, and S. Miles. The rationale of PROV. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35, Part 4:235 – 257, 2015.

[Mor11] L. Moreau. Provenance-based reproducibility in the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):202 – 221, 2011.

[Tuf13] Z. Tufekci. Big data: Pitfalls, methods and concepts for an emergent field. Technical report, March 2013.

[10] `http://www.qsrinternational.com/`