Data and Services Discovery projects – Transformative Data Collections

Title

National Australian Twitter Collection

Approach

The ARDC Discovery project to develop a National Australian Twitter Collection (NATC) seeks to extend and enhance existing frameworks for gathering large-scale data from the Australian Twittersphere into a truly national resource that can be made available to all Australian researchers. It builds on prior research and development efforts led by the QUT Digital Media Research Centre (DMRC) through the 2014-16 ARC LIEF project TrISMA: Tracking Infrastructure for Social Media in Australia (LE140100148), in partnership with Curtin, Deakin and Swinburne universities and the National Library of Australia, and on subsequent development efforts at the DMRC and QUT Digital Observatory. This has established a collection that contains some 3 billion tweets by up to 3.72 million Australian Twitter accounts between the launch of Twitter in March 2006 and the present time, and is growing by some 8-900,000 tweets per day at present.

The NATC project builds on this prior work and begins the process of transitioning it to a fully national, sustainable and continuing effort. The project is led by the QUT DMRC, and supported by Curtin University, Deakin University, Monash University, RMIT University, Swinburne University of Technology, University of Adelaide, University of Canberra, University of Melbourne, the University of Queensland, and University of Tasmania. Prof. Axel Bruns is the project leader.

The National Australian Twitter Collection is envisaged as a continuous, real-time collection of Australian tweeting activity that can be used to observe current, short-term events as well as continuing longitudinal developments. The NATC is a globally unique asset: a similar nationally specific collection of Twitter content of this scale is not available for any other nation in the world, and it fills a critical gap in our understanding of specific local and regional use practices for social media platforms. For ethical and privacy reasons this asset cannot simply be made available publicly; rather, we aim for it to be accessible to researchers at Australian universities who can demonstrate ethical clearance from their home institution for their intended research activities, and who have received NATC accreditation on that basis. We also intend to deposit the NATC dataset with the National Library of Australia for long-term preservation and usage.

The ARDC Discovery project represents a prototyping stage towards these goals. Building on the prior work conducted by QUT and its partners, it focusses on three central and interrelated elements: 1) continued development of the NATC data gathering, processing, and storage architecture; 2) recruitment of a sector-wide alliance of research and research services institutions supporting the NATC initiative; and 3) identification of legally and regulatorily sound institutional frameworks for a future NATC entity. Project activities and expenditure were distributed across these three elements.

It is important to highlight in this context that the NATC dataset is a continuous, live dataset that accumulates Australian Twitter activity data from the launch of Twitter in March 2006 into the foreseeable future. This means that the ARDC Discovery project itself does not simply produce a complete and final dataset, but instead constitutes a stage in a longer-term effort that focusses on establishing the NATC as a permanent institution providing Twitter data access to all Australian researchers – an institution that may also serve as a template for subsequent initiatives providing data on social media platforms other than Twitter. The establishment of this institution will constitute a truly transformative moment in Australian and international social media research.

As a result of institutional concerns about the contracts initially provided by Monash University acting as agent on behalf of the ARDC, the timeline for the present ARDC Discovery project has been foreshortened considerably. QUT's legal concerns about the initial terms stipulated by Monash, and subsequent adjustments to the contract text, delayed sign-off on the project (and thus access to project funding) from mid-June to mid-August, leaving less than two months for work on the project. While we have managed to produce significant outcomes despite these challenges, we suggest that the ARDC take the necessary measures to avoid such delays in future rounds of this project scheme.

Our three strands of activity over the course of this project are as follows:

1) Platform Development. The NATC builds on existing data gathering methods and tools developed by the TrISMA ARC LIEF project and QUT Digital Observatory, using NeCTAR infrastructure for data gathering and Google BigQuery for data storage, processing, and access provision. Especially because of the increasing unreliability of NeCTAR as an infrastructure for continuous, live, mission-critical data gathering, this mix of technologies is unlikely to represent a feasible foundation for future stages of the NATC. The project therefore committed some of its budget to the translation of existing methods and tools for the Amazon Web Services (AWS) platform, with further development and financial support provided by QUT's Office of eResearch. A prototype of the NATC software is now functional on this platform, and over time will also enable the more precise estimation of future running costs for the NATC.

2) Partner Recruitment. The NATC ARDC Discovery project is led by QUT and supported by ten other leading Australian universities. The project held a meeting of key representatives from these universities in Brisbane on 30 September – as a satellite event to the 20th conference of the international Association of Internet Researchers, hosted by QUT on 2-5 October. The meeting provided an opportunity for participants to review the current state of progress in the project,

reconfirm their institutions' commitment to the NATC initiative, and most crucially also to deliberate on further pathways for the establishment of a national alliance supporting the NATC and its transformation into a persistent entity. Building on the outcomes of these deliberations, we have also further engaged with the National Library (NLA) and Australian Academy of the Humanities (AAH), and will seek to add additional key stakeholders to this alliance as the NATC moves to its next phase beyond the ARDC Discovery project itself.

3) Legal Frameworks. This next stage of the NATC development also depends crucially on the development of appropriate institutional frameworks for a future national entity charged with operating the NATC; this has become increasingly critical as changing regulatory environments and platform rules have affected social media data gathering efforts, for Twitter and other social media services, in Australia and elsewhere. To complement and support the major activities outlined in the original project outline, we therefore decided to direct a significant component of the project budget to seeking sound and up-to-date legal advice from technology law specialists King & Wood Mallesons (KWM), focussing especially on the legality of the NATC's data gathering efforts under Twitter's terms of service and Australia's data and privacy regulations, and on the most appropriate institutional frameworks for a NATC entity operating the collection on a continued basis, either as a separate institution or under the auspices of a university, sector-wide body like the ARDC, or statutory national institution like the National Library or National Archives. This legal advice affirms the overall approach taken by the NATC project, and outlines several options for the future operation of the NATC. It is currently in draft, and will be made available as an additional attachment to this report once it has been finalised.

FAIR

Please see the attached FAIR assessment spreadsheet.

Collaboration and coverage

The ARDC Discovery project has enabled us to expand on the existing consortium of partners in the TrISMA ARC LIEF project and its post-LIEF extensions, and to establish an alliance of supporting institutions that now includes ten Australian universities in addition to QUT. We are exploring further opportunities for engaging with the National Library, the Australian Academy of the Humanities, and other relevant national bodies; Prof. Bruns will speak about the project at AAH-hosted events in Canberra on 17 and 18 October and will also engage further with the NLA at this time. This work prepares the ground for the next stage of the NATC initiative.

The legal advice received from KWM similarly provides the foundation for the formalisation of existing Australian Twittersphere data gathering frameworks in an ongoing national entity that adopts best-practice data governance frameworks under Australian law.

The content gathered and made available by TrISMA and subsequent iterations of this initiative has always been national in its coverage, and this aspect positions the NATC as a globally unique undertaking; there are no similar national social media datasets anywhere else in the world. This

provides unique research opportunities and a significant competitive advantage for Australian researchers. New NATC enhancements prototyped by this project will sustain this advantage.

Sustainability

In light of the continuous, live nature of the social media data gathered and made available by the NATC, sustainability remains its most critical challenge. The ARDC Discovery project is a step towards long-term sustainability: we have established the basis for a truly national user community by building a considerably larger alliance of supporting institutions; are in negotiations to enrol the NLA and AAH as future partners; are planning further grant applications to support the NATC; and have commissioned legal advice (currently in draft) on the most appropriate institutional frameworks for a future NATC entity at the national level. QUT also continues to sustain current data collection efforts through significant financial and in-kind support provided through its Office of eResearch, Digital Observatory, and Digital Media Research Centre.

The next steps in the development of the NATC are likely to be organisational rather than technical: in the coming months we will engage in sustained lobbying with relevant entities at the national level to support the establishment of the NATC as a formal entity, either within relevant existing organisations such as the NLA or ARDC or as a stand-alone organisation, for the longer term. This will also require the further concretisation of clear accreditation and data access frameworks that are available to university researchers throughout the country.

Learnings

Our project is likely to be somewhat unusual in the context of the ARDC Discovery scheme as it is concerned with establishing an ongoing, live data collection and access initiative, rather than only with developing or transforming existing, complete, or closed datasets. It is crucial to understand the work of this project as a stage in the longer-term effort to transition the outcomes of the TrISMA project and subsequent post-ARC LIEF efforts by the QUT Digital Observatory into a continuing and sustainable national entity available to the entire sector. There is no single, final dataset that is made available at the conclusion of the ARDC Discovery project; rather, through technical enhancements, partner recruitments, and legal advice we have made considerable progress towards the long-term goal of establishing the National Australian Twitter Collection.

The ARDC may need to further adjust its Discovery documents, requirements, and processes to anticipate projects of this kind, as especially with the computational turn in the digital humanities they are likely to become increasingly common. Again, we also strongly suggest that it revisit the default contracts offered to project grantees, as it is severely disruptive to projects that are already funded only for a short duration if they lose another two months from being stuck in legal limbo.

Impact

Social media platforms serve as critically important data sources on contemporary public debate, current events, and the continuing evolution of public opinion. They enable the observation of

societal dynamics at an unprecedented level of detail and in close to real time, and provide insights both on acute, short-term, current events and on the longitudinal, gradual change in views, attitudes, and communicative practices over months, years, and decades. Additionally, social media platforms and their affordances are themselves implicated in important social phenomena – for instance, current debates about 'fake news' and other forms of disinformation – and platform providers' action or inaction on these matters directly impact on society. For all these reasons, social media data provide an excellent indicator of societal dynamics.

However, for technical and practical reasons the gathering of data from social media often remains episodic, centred around key moments (elections, crises, events) and elite actors (politicians, journalists, celebrities), with a significant international rather than Australian focus. The National Australian Twitter Collection is a truly unique resource: internationally, it is the only long-term comprehensive collection of social media activity on a single platform at a national scale, and enables unprecedented insights into current political developments, societal issues, and everyday life in Australia. Such insights are valuable both now, represented by uses of the NATC as a live resource, and well into the future, as the NATC collection becomes a unique and in-depth historical archive of social media activities in Australia since the early 2000s.

The TrISMA dataset and its subsequent iterations have already been used widely to enable research into a wide variety of issues, topics, and events. By developing sustainable technological and institutional frameworks for the continuing operation of the NATC, this ARDC Discovery project has identified workable pathways towards the long-term continuation of such research efforts, and towards the extension of access to these datasets, under appropriate accreditation models, to researchers at all Australian universities. This further cements the leading role that Australian social media (and social media-enabled) research already plays at the global level.

Report prepared by:
Prof. Axel Bruns, Digital Media Research Centre, Queensland University of Technology
Date: 8 Oct. 2019