

Building an Australian Social Data Observatory

A national system to analyse and make use of the data in our ever-expanding digital world

BY JULIAN THOMAS, JEAN BURGESS, DANIEL ANGUS AND AMANDA LAWRENCE



Prof Julian Thomas is Director of the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S Centre), and a Distinguished Professor in the School of Media and Communication at RMIT University. He also leads the team producing the Australian Digital Inclusion Index since 2015. His work ranges across the contemporary histories of new communications technologies, digital inequality and inclusion, and the internet and communication policy. Co-authors: Prof Jean Burgess, Prof Daniel Angus and Dr Amanda Lawrence

I

IN MARCH 2021, FACEBOOK announced it was building Instagram for under-13s¹. By May, 44 US Attorneys General had written to Mark Zuckerberg demanding that he stop the project because, amongst various concerns, 'Facebook has historically failed to protect the welfare of children on its platforms'.²

Despite his assurances that children would be protected on the new platform, a year later it was revealed that Facebook doesn't actually know where all of its data goes.³

Digital platforms drive our social and economic life, delivering educational programs and government services, mediating democratic debate and decision-making, selling products, connecting communities, families and friends, distributing knowledge and recommending media, entertainment and information options for billions of people worldwide.

The potential benefits to the Australian economy through digitalisation are estimated to be as much as \$315 billion over the next decade, with the potential to create up to a quarter of a million new jobs by 2025.⁴

Of course, online platforms, including social media services, may be used for

both harmful and beneficial ends. The question is, how do we know which is happening? How do we increase the benefits and reduce the risks of these critical technologies and the data they generate?

Opening up the 'black box' of digital platforms and their algorithms is essential as we move to an increasingly digitised economy and society.

It is essential for effective regulation and legislation to protect and enhance the digital experience for consumers, businesses and the community, just as it is to ensure we have responsible, ethical and inclusive online spaces for all, including our children.

New sources of data, such as mobile and social data, have allowed researchers to model and understand the circulation of information, goods and services in ways that were not possible just a decade ago.

There has been a corresponding explosion in the development of tools and techniques for the analysis of these 'digital traces', their consequences, and the algorithms and platforms that generate them.

Many tools take advantage of advanced computational research methods such as artificial intelligence (AI), data analytics, natural language processing (NLP), image recognition and blockchain.

Despite these advances, our capacities to analyse online data are still limited as a vast amount of data is closed off in corporations' proprietary archives and accessing large-scale social media and other digital data through commercial platforms continues to be a challenge.

The major digital platforms have not historically demonstrated a consistent commitment to public research applications.

While there are some attempts to share data, the application programming interfaces (APIs) provided are often overly restrictive and many platforms have developed sophisticated technical measures to detect and prevent third parties from accessing information in the public domain at the scale that is required for systemic analysis.

In response, researchers have found ways to not only work with major platform companies, using their APIs or data-sharing initiatives, but also work around them, scraping data directly or partnering with users through data donation projects to collect real-time or existing digital trace data.

Partnering with users through data donation programs and crowdsourcing platforms has gained momentum due to the expansion of users' access and portability rights, such as the EU's *General Data Protection Regulation* (GDPR).

Australian researchers are at the forefront in experimenting with these new methods for studying digital platforms.

The Australian Research Council (ARC) Centre of Excellence for Automated Decision-Making + Society (ADM+S) currently has

two data donation projects underway: the Australian Search Experience⁵ and the Facebook Ad Observatory.⁶

However, like many other digital research projects, the tools created are bespoke and do not scale for national or cross-sector applications.

Building on the work of world-leading Australian researchers, we propose building an Australian Social Data Observatory (ASDO),⁷ a national research facility that would provide access to large-scale social, economic and cultural data and the analytical tools and governance required to support cutting-edge research across a range of social, economic, health and environmental issues.

Building ASDO involves four key elements: social data sourcing, a data laboratory, data classification and linking, and governance, training and engagement.

The concept draws on a suite of innovative 'critical simulation' methods for collecting and analysing data from digital platforms, such as data donations, synthetic data and crowdsourcing tools, combined with machine learning, natural language, image processing and other technologies.⁸

Data and analysis could be further enhanced and integrated with other data and tools through linking and standardised labelling. The facility would be driven by best practice governance and ethics for data collection and management using existing protocols and new insights.

The data donation approach may involve software or plugins that monitor user behaviour and collect real-time data such as pages visited in a browser, or apps used on a smartphone, or it can use existing data provided through data download packages (DDPs).

Mobile data, sensors and wearables provide other sources of real-time data that can be donated. Researchers can combine this with data from surveys, interviews and digital ethnography to compliment

and explain large scale social data.

Synthetic data generation is another approach which can be used to develop and test data science algorithms and models without compromising legal rights or research ethics standards (when used appropriately).

The data laboratory would provide a range of machine learning, network analysis and other technologies and approaches for analysing this data.

These include data enrichment approaches such as DeepSpeech, an open-source speech to text library for the automatic transcription of audio and video data; the ImageMachine, a machine vision framework developed by ADM+S researchers to assist in the classification, clustering and sorting of image data; NLP for tagging, topic modelling, and named-entity extraction; and network analysis using dynamic graph databases.

These computational approaches would be enhanced through the development of a crowdsourcing platform to provide access to collective intelligence through citizen science and community collaborations.

ASDO would be a platform for Australian researchers across all sectors to conduct unique crowdsourcing social data projects that engage and work with Australian communities, including Indigenous groups and those who speak languages other than English, to collect, annotate and

process text, images, video and audio data.

Crowdsourcing tools can also be used to annotate and enrich data and combined with machine learning can augment and accelerate analysis of diverse or complex data such as historical documents, images, video and audio.

ASDO would support the input of 'messy' data to be converted into more robust, archival-quality, open data standards for linking with other data sources and tools.

Of course, governance of social data is complex, involving issues of privacy, security, intellectual property, commercial control, bias in data and algorithms, ethics, noise and inaccuracies.

Access and licensing arrangements for 'big social data' generated by proprietary social media platforms are also needed. The problem of intellectual property rights when dealing with commercial data has also not been tackled in any meaningful way.⁹

There are also various challenges for the ethical conduct of social and health data donation research including participant protection, representativeness, incentives to participate and governance.

ASDO would provide an opportunity to explore and test the many ethical, legal and standardisation issues presented by social data research and develop national and international guidelines for best practice.

The idea for ASDO is driven by the needs of researchers at two new Centres of Excellence – ADM+S, and the Centre for the Digital Child¹⁰ – and is backed by university and industry partners from across Australia as well as international collaborators.

The facility would also build on and extend existing research infrastructure investments and collaborate with new facilities such as the Humanities Arts and Social Sciences (HASS) Research Data Commons.

As well as government and institutional investments, it would also be able to generate revenue from services, training and access to facilities based on a not-for-profit business model.

Data donation projects have been running in the health domain for some time as a way of accessing patient data as well as data from those not in the health system, such as through health tracking apps.¹¹ Examples include The Data Science Platform (DSP)¹² developed by the Broad Institute in the US and Data4Life¹³ based in Germany.

Social data donation projects are more recent and include DataSkop from Algorithm Watch¹⁴, Mozilla Rally from the Mozilla Foundation¹⁵, the Citizen Browser developed by independent journalists at *The Markup*¹⁶ in the US, and the recently funded Digital Data Donation Infrastructure (D3I) being developed by six universities in the Netherlands. ASDO would put Australia on the map as a world-leading facility in this emerging field.

By providing the tools and resources to gather and analyse online user experience data, ASDO would dramatically extend access to social media data beyond the small group of specialists who currently work in the field.

Such a capability would enable Australian researchers across HASS and science, technology, engineering and mathematics (STEM), as well as government, industry and civil society to benefit from the insights derived from social data.

The demand for digitally skilled workers is expected to increase by 100,000 between 2018 and 2024. At the same time, emerging technologies such as AI and automation are disrupting the workplace—eliminating, creating or reconstructing jobs—with estimates that 25 to 46 per cent of existing jobs could be automated by 2030.

As a dedicated national facility, ASDO would provide the skills and training needed for a new generation of researchers across all disciplines to use and analyse social data, providing real-time understanding of diverse contexts and complex social, cultural and economic issues.

ASDO aligns with many of Australia's national priorities, including maximising jobs and opportunities from the Digital Economy Strategy and the AI Roadmap.

It also aligns with a raft of policies and strategies including the *Australian Data Strategy*, the *National Data Availability and Transparency Bill*, the *Australian Cyber Security Strategy*, the *Digital Government Strategy 2020*, the *Consumer Data Right*, the long-awaited review of the *Privacy Act (1988)* and the *ACCC Digital Platform Services Inquiry 2020-25*.

As the *2021 National Research Infrastructure Roadmap*¹⁷ highlighted, social data analysis and digital research capabilities are priority areas for a range of disciplines and sectors. ASDO would provide a cross-cutting facility with wide benefits for HASS and STEM as well as government, industry and civil society.¹⁸

It is in our national interest to build the systems we need to understand the world in which we live—not just earth and space, but an ever-expanding digital world, with many galaxies, wonders and black holes. For this, we need the ASDO. ■

We propose building an Australian Social Data Observatory (ASDO), a national research facility that would provide access to large-scale social, economic and cultural data and the analytical tools and governance required to support cutting-edge research across a range of social, economic, health and environmental issues

