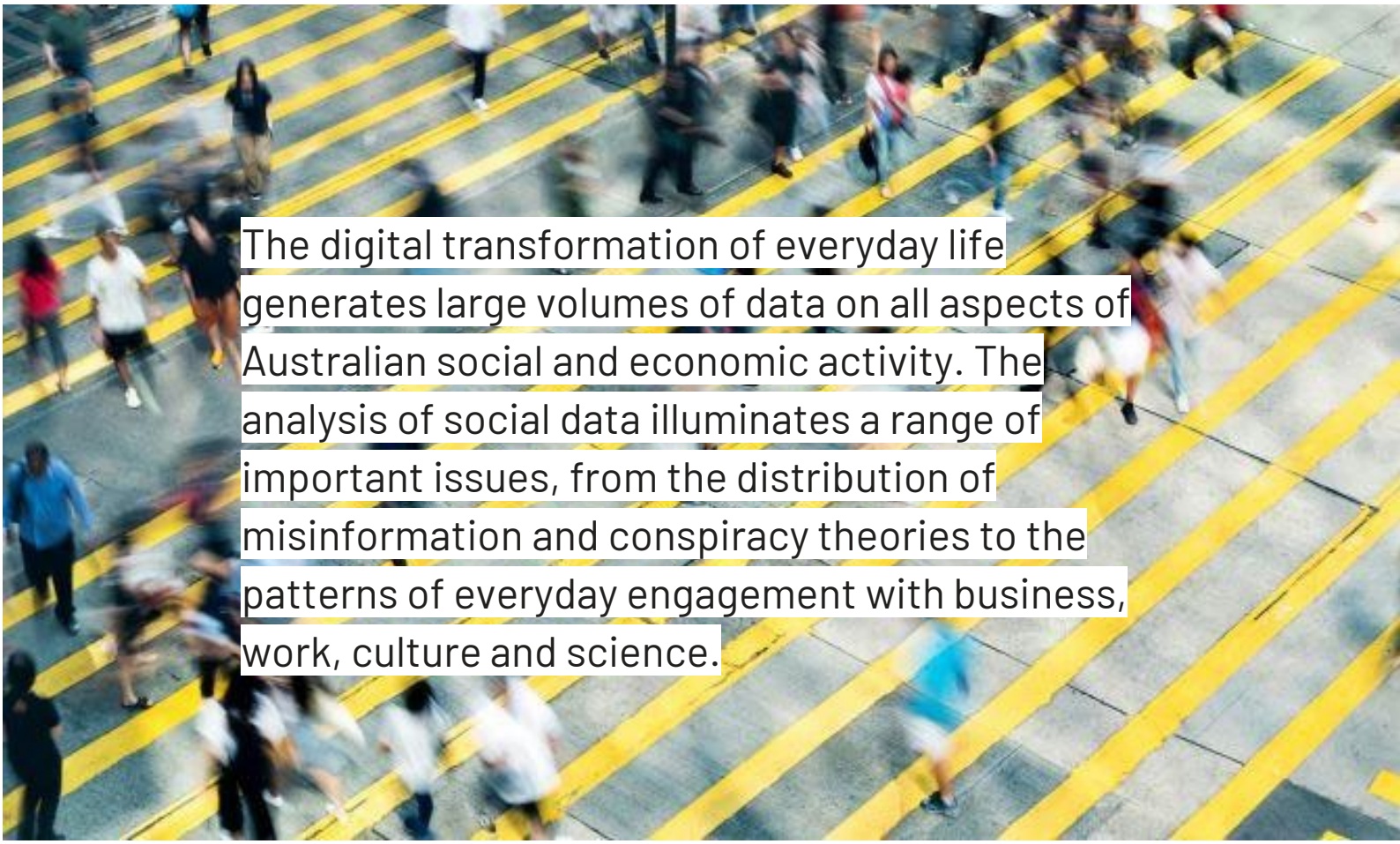
Landmark Social Data Research Infrastructure: Australian Social Data Observatory

Concept Brief

ARC Centre of Excellence for Automated Decision-Making + Society

December 2021
https://www.admscentre.org.au/publications/

ADM +S
ARC Centre of Excellence for Automated Decision-Making and Society



The challenge

New sources of data, such as mobile and social data allow researchers to model and understand the circulation of information 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, data analytics, natural language processing, image recognition and blockchain.

However, despite the vast quantities of social data now circulating, our capacities to analyse such data are still very limited. Some constraints relate to the nature of the data involved: and increasingly access may be subject to the policies and licensing arrangements of private entities. The major digital platforms have not historically demonstrated a consistent commitment to public research applications. Public data may raise different issues, e.g. the application of research ethics standards. Other constraints on social data research have to do with the limited knowledge and skills in data analysis among researchers in Australia.

Digital technologies are expected to contribute $65 billion to Australia's GDP by 2023, however without dedicated investment in social data research infrastructure we risk falling behind in research and innovation, and successful application of emerging technologies. Social data analysis is essential for research across a range of disciplines and sectors to ensure we are prepared for our future as a digital economy and society.

The Australian Social Data Observatory (AuSDO)

The Australian Social Data Observatory (AuSDO) is a proposal to develop landmark National Research Infrastructure (NRI) for social data to support research across a range of disciplines and national priorities including advancing the digital economy, future manufacturing, recycling and clean energy, transport and health. A national facility such as AuSDO will provide tools and capabilities to gather and analyse online user experience data, making social data dramatically more useable for Australian researchers across all sectors.

AuSDO builds on current investments and existing facilities across the NRI sector where possible, however this proposal is a significant advance on what currently exists at the institutional or national level. Given the difficulties in accessing large-scale social data through commercial platforms, AuSDO will provide tools and guidelines for cutting edge approaches such as data donation, crowdsourcing and citizen science, allowing researchers to access and analyse digital data from multiple platforms. Specialist knowledge of the platforms will no longer be required, nor the development of ad hoc analytic tools.

AuSDO involves four key elements: Social data sourcing, Data tools and applications, Data classification and linking, and Governance, training and engagement, as illustrated below.

Social Data Sources

AuSDO will assist researchers in accessing and analysing a range of social data which may be derived directly from platforms, either through official or unofficial Application Programming Interfaces (APIs), brokered through connections to existing platforms such as the Australian Digital Observatory, or directly from users in the form of crowdsourced or donated data. It will also offer integrations with existing datasets, including traditional social statistical information, hosted on other national and international platforms such as the HASS Research Data Commons, the Population Health Research Network (PHRN), and the Australian Urban Research Infrastructure Network (AURIN) – as illustrated below.

> Data donations provide an effective means for researching digital platforms and algorithmic systems without having direct access to them. Users can donate their data individually, for example when using social media platforms, apps, credit scoring services or when shopping online. In this way it is possible to shed light on the black-box algorithms of these systems and to understand how they generate recommendations, evaluations and content-related decisions. ADM+S is already working on data donation projects with the Australian Search Experience and the Australian Ad Observatory which will provide the knowledge required to expand this approach within AuSDO to enable more researchers across multiple disciplines access to tools and training to apply these methods.

> Crowdsourcing is a means of collecting and analysing data using collective intelligence or micro tasks. This is widely used in commercial systems such as Mechanical Turk and Figure Eight however it can be extremely effective for public interest research through open source citizen science platforms such as Zooniverse, iNaturalist, Wikidata, Pybossa and others. AuSDO will provide best practice guidelines and tools to enable Australian researchers to conduct crowdsourcing social data projects to annotate and process text, images, video and audio data.

> Web scraping is a commonly used technique for gathering data from web platforms where no official API or other data download mechanism exists. Web scraping is useful in tracking changes to government websites, news and media content, public forums, and commercial platforms (e.g. tracking changes to published Terms of Service). Google Trends and search results from digital platforms will also be a key source of data. Researchers require the tools and training to apply these methods in their projects.

> Synthetic data generation is a relatively new approach which may be used to develop and test data science algorithms and models without compromising legal rights or research ethics standards (when used appropriately). Tools include open source systems such as the Synthetic Data Vault, recently developed by the Data to AI Lab at MIT which allows synthetic data to be generated in place of real data, or in addition to real data, as an enhancement. Providing this technology in AuSDO will enable researchers in Australia to conduct ground-breaking studies in a variety of areas such as public health, behavioural science, algorithmic bias and recommendation systems.

AuSDO integration with the NRI Ecosystem

Social Data Laboratory: software, tools and applications

AuSDO will provide an integrated laboratory of software services, tools and applications for researchers to learn and apply across diverse social data sources. These will be supported through enabling systems such as cloud computing and secure access.

> **DeepSpeech,** an open source speech to text library supported through the Mozilla foundation can be utilised within AuSDO to assist researchers in the automatic transcription of audio and video data.

> **NLTK**, the Natural Language ToolKit, is an open source NLP library that can assist in enriching text-based data within AuSDO, including functions such as part-of-speech tagging, topic modelling, and named-entity extraction.

> **Raphtory**, an open source dynamic graph developed by the UK's Turing Institute, is a cutting-edge approach to graph databases which would be deployed in AuSDO to analyse data (either previously stored, or a real-time stream), to create dynamic graphs showing connections and change over time. This is particularly useful for social network analysis.

> **ImageMachine**, is a machine vision framework developed by ADM+S researchers that would be made available in AuSDO to assist in the classification, clustering and sorting of image 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.

> **Enabling systems such as cloud computing and secure access** will provide critical infrastructure and computational capacity for AuSDO including the National Computational Infrastructure, NECTAR Research cloud, AARNET and the Australian Access Federation.

Data Classification and Linking

Social data is often 'thick' data, with different data types and formats (pictures, video, text, numerical), multiple fields, variable degrees of structure, and additional contextual meaning attached to any single data record. AuSDO will enable the linking of social data and will also support the input of 'messy' data to be converted into more robust, archival-quality, open data standards.

Governance, Training and Engagement

The Australian Social Data Observatory will be managed as a national facility, with a management structure reflecting a broad range of stakeholders beyond ADM+S. It will be developed in close consultation with the Australian Academy of the Humanities and through engagement and collaboration with existing NCRIS and ARDC facilities.

Governance of social data is complex, involving issues of privacy, security, intellectual property, commercial control, bias in data and algorithms, ethics, noise and inaccuracies. AuSDO will develop and implement guidelines for managing data based on adoption and implementation of the Findable, Accessible, Interoperable and Reusable (FAIR) principles and the Collective Benefit, Authority to control, Responsibility and Ethics (CARE) principles for Indigenous data governance and integrate them into all aspects of the facility. Social data analysis is a relatively new field of data science and AuSDO will build a world leading research facility generating new approaches to data management, ethics and techniques.

AuSDO will provide training, tools and guides for Australian researchers across HASS and STEM disciplines including media and communications, law, economics, politics, history, sociology, the behavioural sciences, public health, environmental sciences, manufacturing and recycling. It will also provide extensive opportunities for engagement and collaboration with government, industry and civil society organisations increasingly dependent on understanding social data.

Benefits

The Australian Social Data Observatory will be a world leading research facility for the Australian community, providing the cross-cutting capability to support collaboration and data intensive analysis and modelling on critical social, economic and public interest issues using big data from digital platforms and other sources. AuSDO will provide integrated access to diverse and large-scale social data and the tools, technologies and governance required for efficient and cost effective analysis and research impact across many disciplines.

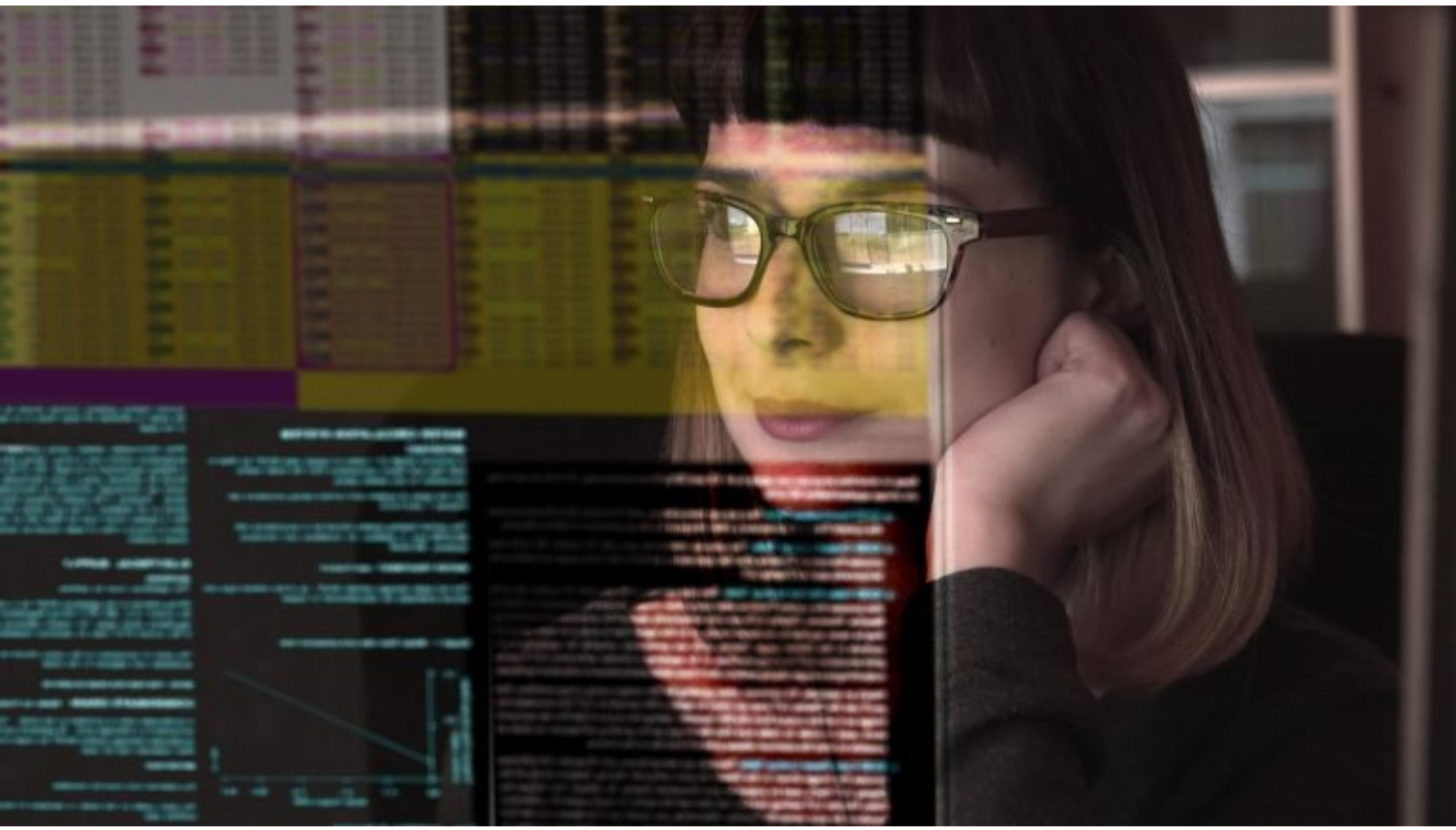
A social data observatory aligns with national government priority frameworks including the Digital Economy Strategy, Australian Data Strategy, Blueprint for Critical Technologies (opportunities for improved health and social outcomes), the National Climate Resilience and

Adaptation Strategy (social/health domain), Artificial Intelligence Roadmap and the National Science and Research Priorities (manufacturing, health and transport).

Many of the challenges identified in the 2021 NRI Roadmap Exposure Draft highlight the need for social and human data and human-centred design including for novel medical products and processes, innovative recycling and clean energy initiatives and frontier technologies and manufacturing. A national system for curating and analysing social and human behaviour data from diverse sources would also provide a cost effective response to cross-cutting issues such as the need for system-wide enhancements to NRI including integrated datasets, software analysis tools and platforms, and contribute to cutting-edge national digital research infrastructure (NDRI).

Digital technologies play an increasingly important role across the economy, society, culture and innovation with the demand for digitally skilled workers expected to increase by 100,000 between 2018 and 2024. At the same time emerging technologies such as artificial intelligence and automation are disrupting the workplace—eliminating, creating or reconstructing jobs—with estimates that 25–46 per cent of existing jobs could be automated by 2030. As a dedicated national facility, AuSDO will 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.

The aim is to extend access to social media data beyond the small group of specialists who currently work in the field. AuSDO will enable researchers across HASS and STEM, as well as those in government, industry and civil society, to benefit from the enormous resources available through social data. It will contribute substantially to the digital transformation of both HASS and STEM sectors in Australia.



Consultation

This concept brief is the product of ongoing engagement with social data and research infrastructure projects and facilities by researchers at ADM+S and their institutions and research centres in consultation with the Australian Academy of the Humanities, the ARDC, the HASS RDC and other NCRIS facilities. We look forward to discussing the ideas in this concept brief further with interested stakeholders from across the research and innovation sector.

About ADM+S

The ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S) is a cross-disciplinary national research centre. ADM+S aims to create the knowledge and strategies necessary for responsible, ethical, and inclusive automated decision-making. It brings together leading researchers in the humanities, social and technological sciences in an international industry, research and civil society network. Its priority domains for public engagement are news and media, transport, social services and health. The Centre's four Research Programs — Data, Machines, Institutions, and People — examine all the different elements that constitute automated decision-making systems.

ADM+S is a collaboration between nine Australian universities: RMIT University (host institution), Monash University, Swinburne University, Queensland University of Technology, University of Melbourne, University of New South Wales, University of Queensland, University of Sydney and Western Sydney University. The Centre also partners with eight universities from around the world. Industry and civil society partners include Google, Telstra, Bendigo Health, Australian Red Cross, the ABC, Australian Communications Consumer Action Network (ACCAN), Algorithm Watch, and the Digital Asia Hub.

ADM+S researchers have extensive experience and engagement with initiating and managing research infrastructure facilities as well as expertise in research and innovation policy, governance and business models. ADMS Researchers and partner organisations have been involved with institutional and national research infrastructure facilities including: Analysis & Policy Observatory (APO), Data Co-op (Swinburne), AURIN, Australian Data Archive (ADA), Trisma, the QUT Digital Observatory, Austlii, the Australian Text Analytics Platform and the Language Data Commons of Australia (LDaCA).

This concept brief was prepared by: Prof Julian Thomas, Prof Jean Burgess, Prof Daniel Angus, Prof Sarah Pink, Prof Anthony McCosker, Prof Axel Bruns, Prof Kimberlee Weatherall and Dr Amanda Lawrence. Further advice was provided by Prof Fiona Haines, Prof Christine Parker and Prof Megan Richardson.

www.admscentre.org.au

ADM+S ARC Centre of Excellence for Automated Decision-Making and Society