Australian Social Data Observatory (ASDO)

National Research Infrastructure for the digital economy and society

October 2022

October 2022

ARC Centre of Excellence for Automated Decision-Making + Society

www.admscentre.org.au/asdo

Prepared by the ADM+S National Research Infrastructure Working Group: Prof Julian Thomas, Prof Jean Burgess, Prof Daniel Angus, Prof Anthony McCosker, Dr Dang Nguyen, Dr Damiano Spina and Dr Amanda Lawrence.

Organisations consulted in the development of this proposal include: the ARC Centre of Excellence for the Digital Child, Australian Academy of the Humanities, Australian Social Science Academy, Australian Research Data Commons, HASS Research Data Commons.

ASDO university partners:

RMIT University, Queensland University of Technology, University of Queensland, Curtin University, Swinburne University and Deakin University.

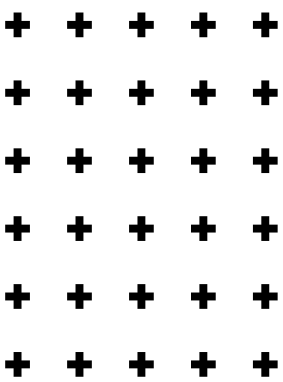
About ADM+S

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.

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.

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.



To reap the benefits and avoid potential harms from the digital transformation occurring across all sectors of Australian society and economy we must invest in the tools and infrastructure required to observe, analyse and design digital platforms.

The challenge

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 next decade will see a new wave of digital transformations and many changes in the way data, communication and digital platforms operate.

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. Building on world-leading research, Australia now has the capability necessary to maximize the economic and social benefits of AI and automated digital services.

Of course online platforms, including search engines, social media platforms, databases and online 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?

Large-scale online data has the potential to substantially contribute to the capacity of business, government and the community to understand and address Australia's major social, economic and public policy challenges. However without dedicated investment in national research infrastructure (NRI) we risk falling behind in research and innovation with emerging technologies.

Opening up the 'black box' of digital platforms and their algorithms is essential for an expanding digital economy and society, for effective regulation and legislation to protect and enhance the digital experience for consumers, businesses and the community, and 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, data analytics, natural language processing, 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 proprietary archives of corporations 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 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).

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. Examples include the The Data Science Platform (DSP) developed by the Broad Institute in the US and Data4Life in Germany. Social data donation projects are more recent and include DataSkop from Algorithm Watch, Mozilla Rally from the Mozilla Foundation, the Citizen Browser developed by independent journalists at The MarkUp in the US, and the recently funded Digital Data Donation Infrastructure (D3I) being developed by a consortia of six universities in the Netherlands.

Australian researchers are at the forefront in experimenting with these new research methods for studying digital platforms. ADM+S currently has two data donation projects underway–the Australian Search Experience and the Facebook Ad Observatory. However like many other digital research projects, the tools created are bespoke and do not scale for national or cross-sector applications.

There is an enormous opportunity to invest in national infrastructure and the skills and expertise required to access and analyse social data from multiple platforms in innovative ways to maximise the benefits for the digital economy and society over the next decade and beyond. However Australia's NRI framework has not yet responded to the major challenge of collecting, analysing and connecting the national social data and analytical tools required to support research on the social, cultural and economic benefits and challenges of digital transformation.

The ADM+S Australian Ad Observatory Project is investigating targeted advertising on social media using data donations.

The *2021 National Research Infrastructure Roadmap* identified social data analysis and digital research capabilities as critical areas of investment for research across a range of disciplines and sectors to ensure we are prepared for our future as a digital economy and society. To address these challenges we propose establishing an Australian Social Data Observatory as a new, cross-cutting, national level facility.

The Australian Social Data Observatory (ASDO)

The Australian Social Data Observatory (ASDO) is a proposal to develop landmark National Research Infrastructure 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, food and beverage, transport and health.

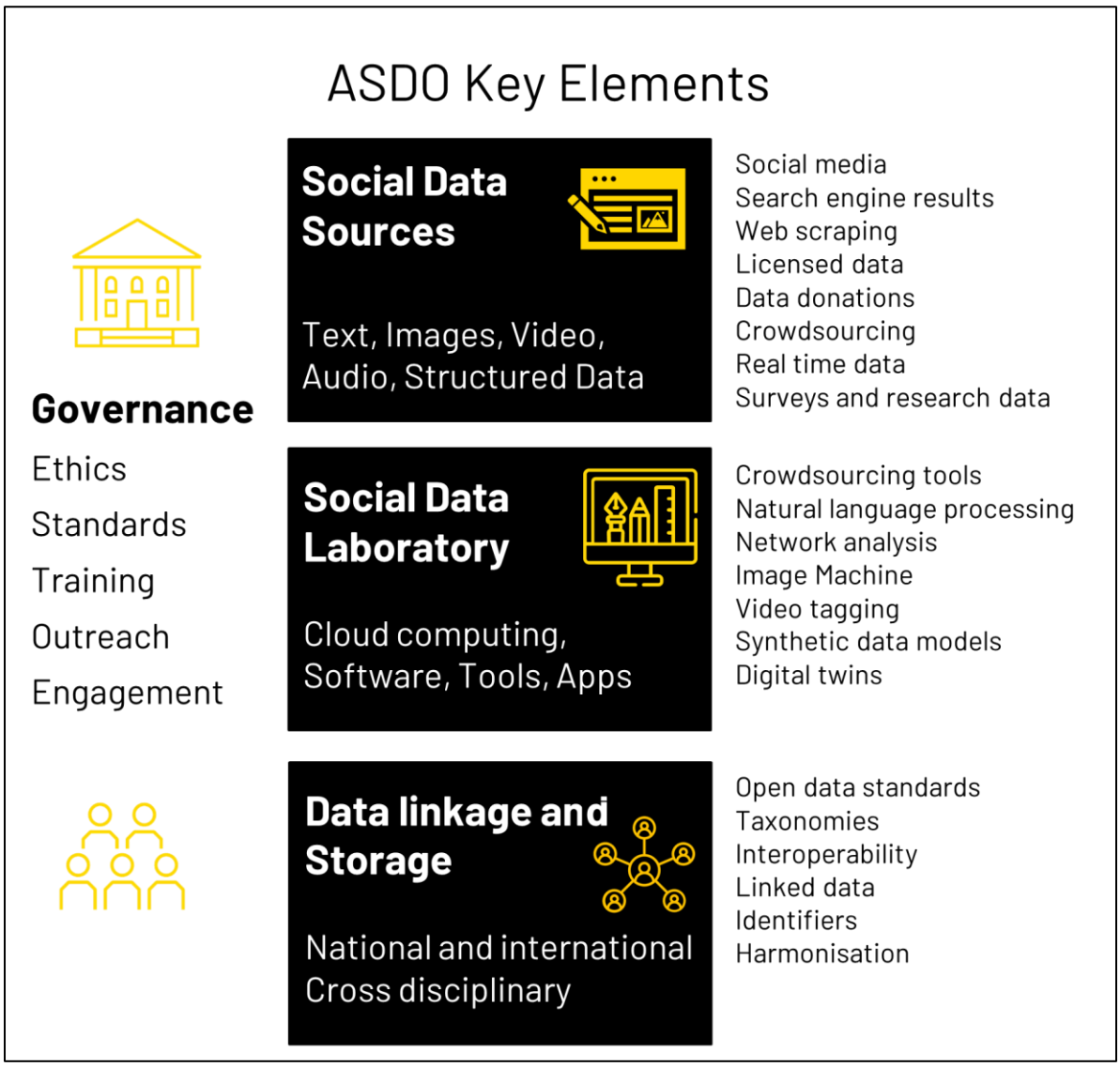
Social data and data analytic tools would enable research on critical national issues from the distribution of misinformation to the patterns of everyday engagement with business, culture and science, the flows of communication in emergencies and humanitarian crises, and the dynamics of political conflict and consensus.

A national facility such as ASDO will provide tools and capabilities to gather and analyse online user experience data, algorithms and interactions, making social data dramatically more useable for Australian researchers across all sectors. ASDO involves four key elements: Social Data Sourcing, a Social Data Analytics Platform, Data Linking and Governance.

Social Data Sources

ASDO 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).

> 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. Google Trends and search results from digital platforms will also be a key source of data. Data donations are a way 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 experience with data donation projects will provide the knowledge required to expand this approach within ASDO 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 Appen, as well as open-source citizen science platforms such as Zooniverse, iNaturalist, Wikidata and Pybossa. ASDO will establish a platform and guidelines to enable Australian researchers to conduct crowdsourcing social data projects to annotate and process text, images, video and audio data and support pipelines into external systems as required.

> Web scraping is an approach commonly used 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 Terms of Service).

> 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, or in addition to real data. Providing this technology in ASDO 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.

Social Data Analytics Platform

ASDO 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 ASDO 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 ASDO, including functions such as part-of-speech tagging, topic modelling, and named-entity extraction.

> Image Machine: an image vision framework will assist in the classification, clustering and sorting of image data. A prototype of this system is currently being used for the ADM+S Ad Observatory project. Image data is fed into an image analysis pipeline, which implements numerous techniques that enhance the ability to identify content ahead of manual qualitative analysis via the crowdsourcing tools (Burgess et al. 2022). The Image Machine will include more open-ended machine vision approaches to perform clustering and classification of ads based on latent visual properties, not just based on specific objects and text.

> 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 ASDO including the National Computational Infrastructure, NECTAR Research cloud, Amazon Web Services and Google Cloud Console . Access will be supported via AARNET and the Australian Access Federation.

Data Classification and Linking

Social data is often 'thick' data, with different types and formats (pictures, video, text, numerical), multiple fields, variable degrees of structure, and additional contextual meaning attached to any single data record. ASDO 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.

The project will contribute to harmonisation across heterogeneous dataset and develop new metadata standards and protocols to increase efficiency and usability of social data. Work already underway within the HASS RDC will be critical for this process.

Governance, Training and Engagement

Governance of social data is complex, involving issues of privacy, security, intellectual property, commercial control, bias in data and algorithms, ethics, noise and inaccuracies. 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. ASDO 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 ASDO will build a world leading research facility generating new approaches to data management, ethics and techniques.

ASDO has been developed in consultation with a wide range of experts and organisations from across HASS and STEM disciplines including the ARC Centre of Excellence for the Digital Child, and other infrastructure projects, particularly the HASS Research Data Commons and the Australian Research Data Commons (ARDC). It will be developed in partnership with the ARDC and the HASS RDC and supported by six university partners, RMIT, QUT, UQ, Curtin, Swinburne, and Deakin. The project will be overseen by a Steering Group and Stakeholder advisory group with representatives from a broad range of disciplines and sectors. It will also be developed through engagement and collaboration with government, industry and civil society organisations.

ASDO 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.

> *ASDO would provide tools and resources to deliberately gather and analyse online user experience data, dramatically extending access to social media data beyond the small group of specialists who currently work in the field…Such a capability would enable 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*.
>
> 2021 National Research Infrastructure Roadmap

Connections with existing NRI

ASDO builds on and enhances the investments made by NCRIS, the ARDC and the Australian Research Council (ARC) in infrastructure projects for social science and humanities research and will provide integrated capabilities, training, tools and data to support critical social and economic research across a range of disciplines and sectors. ASDO will be developed in close collaboration and integration with the HASS and Indigenous Research Data Commons–IRISS, LDACA, Trove and the Indigenous RDC. In addition ASDO will work closely with existing tools and services including the Australian Digital Observatory (ADO) providing an API for social media platforms data harvesting, the Australian Text Analytics Platform (ATAP), the Virtual Observatory for the Study of Online Networks (VOSON) Lab, CADRE for sensitive data and many others.

A social data observatory will also compliment and expand the capabilities of existing NCRIS facilities including the Public Health Research Network (PHRN), the Australian Urban Research Infrastructure Network (AURIN), the Atlas of Living Australia. It will also provide tools and services that can be integrated with government, civil society and industry platforms and projects.

ASDO will build on current investments and existing facilities across the NRI sector for data and tools where possible, however the proposal here is for 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, ASDO will also provide tools and guidelines for data donation, crowdsourcing and citizen science approaches, allowing researchers to access and analyse social data from multiple platforms.

Benefits

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, ASDO 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.

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 national digital research infrastructure (NDRI). ASDO aligns with many of Australia's national priorities, including maximizing jobs and opportunities from the Digital Economy Strategy and the AI Roadmap. It also aligns with a raft 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*.

The Australian Social Data Observatory will be a world leading research facility for the Australian community, providing the 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. ASDO 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. 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.

ASDO is developed in collaboration with:

RMIT UNIVERSITY

QUT

ADM+S

The University of Queensland AUSTRALIA

Curtin University

ARDC

Australian Research Data Commons

SWINBURNE UNIVERSITY OF TECHNOLOGY

DEAKIN UNIVERSITY

HASS Research Data Commons and Indigenous Research Data Capability