Australian Research Data Commons

Media(ted) Data Focus Area - Australian Internet Observatory

Draft project plan for public feedback

Julian Thomas

08/03/2024

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

ARDC is enabled by NCRIS

CONTENTS

CONTENTS 1
REVISION HISTORY 2
1. PROJECT INFORMATION 2
1.1. Project aims, scope and outcomes 2
The digital transformation challenge 2
Who will this project benefit? 4
Proposed Structure of the Project (in scope): 5
Stream 1. Governance and Data management 6
Stream 2. Community outreach and Research Training 6
Stream 3. Data sourcing 7
Stream 4. Data Laboratory 9
1.2. Budget 10
1.3. Payment Schedule 10
1.4. Project Partners & Subcontractors 10
1.5. Project team roles and responsibilities 13
1.6. Governance 13
1.7. Implementation & Timing - Milestones and Deliverables 14
Stream 1. Governance and Data management 15
Stream 2. Community outreach and Research Training 17
Stream 3. Data sourcing 19
Stream 4. Data Laboratory 23
1.8. Visual Summary Overview 25
1.9. Assumptions 25
1.10. (Inter)dependencies 26
1.11. Risks 27
1.12. Outputs and Outcomes Monitoring and Evaluation Plan 30
1.13. Communications & engagement 34
2. GLOSSARY OF TERMS 35
3. CHANGE CONTROL (for ARDC information only) 35
4. APPENDICES 36

REVISION HISTORY

Version	Date	Editor	Summary of changes
1.0	11 March 2024	Amanda Lawrence	Submission for initial review

1. PROJECT INFORMATION

PROJECT TITLE	**Australia Internet Observatory (AIO)**
PROJECT START AND END DATES	1 July 2024 to 30 June 30 2028
CONTRACTING ORGANISATION	RMIT University
PROJECT LEAD CONTACT PERSON	Prof Julian Thomas
PROJECT MANAGER	Dr Amanda Lawrence
FOCUS AREA and ACTIVITY	Mediated Data, HASS and Indigenous Research Data Commons, ARDC

1.1. Project aims, scope and outcomes

The digital transformation challenge

Digital platforms play a critical role in almost every aspect of Australia's economy and society, yet our capacities to collect and analyse data from digital platforms and observe their activities are still very limited. To address the digital transformation challenge, Australian researchers need new tools and approaches, such as data donations

and simulations, that will provide access to diverse types of digital trace data - text, image, audio, video etc - and new ways of observing and studying online interactions between people, platforms and algorithms.

Such a capability would enable researchers across HASS and STEM, as well as government, industry and civil society, to observe, analyse, understand and respond to the benefits and the dangers of digital platforms. National research infrastructure for mediated data and digital platforms 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.

The Australian Internet Observatory (AIO) is a step-change facility that will support innovative approaches to the collection and analysis of digital social data and internet platforms. It will provide access to large-scale social, economic and cultural data and the analytical tools and governance required to support cutting-edge research on social, economic, health and environmental issues.

By developing a range of new tools and approaches for digital social data and internet research, the Australian Internet Observatory will create an interconnected ecosystem of people, data and tools united through shared technical standards, distributed technical systems, purposefully aligned governance structures and processes, shared open-source tools, and cross-provision of training.

Proposed deliverables (outputs) include:

- Data governance frameworks and implementation plans based on NRI Roadmap and linked data principles
- Ethical and legal frameworks and guides for working with crowdsourcing methods and social data
- A national research training program
- Citizen science data donation program
- An integrated suite of data sourcing and data donation tools including browser extensions, data donation packages and APIs
- Generative AI models for text, audio and image generation
- Test environments and simulation tools
- An integrated suite of open source machine learning tools and data visualisations for analysis and insights

Expected longer-term outcomes include:

Outcomes for the wider community include improvements to informed decision-making and public policy, democratisation and participation in the digital sphere, and public debate, improved digital capabilities and inclusion, greater platform accountability and transparency. The Australian Internet Observatory will contribute to key priorities within NCRIS including:

- **Increased digital skills and capability for researchers across sectors:** The Australian Internet Observatory will provide access to new tools and methods and the skills and training needed for a new generation of researchers across all disciplines to use and analyse social data. This will support and enable a new generation of Australian researchers across both HASS and STEM disciplines to gather and analyse mediated data from a range of digital platforms in new ways.
- **Expand and enhance existing infrastructure and national investments:** The Australian Internet Observatory will build on and expand previously funded research infrastructure such as the Digital Observatory and social media collections at GLAM institutions and other partners. It will also make new social data and digital tools available for integration with existing HASS&I RDC streams as well as the ARDC thematic hubs on health and the environment and NCRIS facilities such as AURIN.

  The Australian Internet Observatory aligns with many of Australia's national research and policy priorities, including the Safe and Responsible AI Review, Combating Misinformation and Disinformation Bill, Digital

Inclusion Strategy, 2021 NRI Roadmap, Digital Economy Strategy, Australian Data Strategy, National Data Availability and Transparency Act (2022), and the ACCC Digital Platform Services Inquiry 2020-25.

- **Innovative technology and national capabilities:** The Australian Internet Observatory 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 approaches, combined with machine learning, natural language, image processing and other technologies. Australian researchers are at the forefront of experimentation with these approaches which are also being adopted by leading research groups in Europe and the US. It will ensure Australian researchers have access to internationally competitive and innovative tools and distributed platforms to help researchers and communities use, annotate, extend and analyse mediated social data.

  A national system for curating and analysing social and human behaviour data from diverse sources will 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).

- **Expanded social data access and collections:** Digital platforms, their architecture, algorithms and content are complex and diverse, comprising quantitative, computation, and qualitative data (including structured numeric and text data, text and documents, images, video and audio). The Australian Internet Observatory will support innovative methods for collecting data from digital platforms, such as through data donations, synthetic data and crowdsourcing approaches. This will be combined with state-of-the-art data science to analyse data at scale including machine learning, generative AI, image processing, emulation environments and other technologies.

- **Collaboration across institutions and sectors:** The Australian Internet Observatory is a multi-institutional partnership operating under a distributed governance model. LIke the rest of the HASS&I RDC, we are committed to a community-driven approach to access and governance of Indigenous data, cultural collections, the need for authentication and authorization of research groups and communities, and the pressing need for tools and skills to enable data-intensive humanities and social science research.

Who will this project benefit?

- University Research Centres and Institutes and Research Organisations including: ARC Centre of Excellence for Automated Decision-Making + Society (ADM+S), ARC Centre of Excellence for the Digital Child, Digital Ethnography Research Centre (RMIT), Digital Cultures & Societies (UQ), Digital Media Research Centre (QUT), Social Innovation Research Institute (Swinburne) and the Centre for Research in Assessment and Digital Learning (Deakin), Alfred Deakin Institute for Citizenship and Globalisation (Deakin), Centre for Resilient and Inclusive Societies (Deakin), Academic Centre of Cyber Security Excellence (UoM), Centre for Artificial Intelligence and Digital Ethics (UoM), Cybernetics Institute (ANU).
- Federal and State Government Departments and Agencies
- Civil society organisations concerned with online safety including: ACCAN, Humanitech, Foundation for Alcohol Research and Education (FARE), community organisations collecting online hate data (e.g., Islamophobia Register Australia, Executive Council of Australian Jewry, Asian Australian Alliance, Call it out)
- Corporates in the technology and internet sectors

- Media organisations
- GLAM Institutions and research Infrastructure facilities and user groups including: National Library of Australia (NLA), NSLA, AURIN, ARDC (HASS&I RDC and People and Planet thematics)

Proposed Structure of the Project (in scope):

The Australian Internet Observatory involves four overarching work streams with various work packages which will run for the duration of the project (illustrated in the figure below):

1. **Governance and Data Management**
2. **Outreach and Research Training**
3. **Data Sourcing**
4. **Data Laboratory**

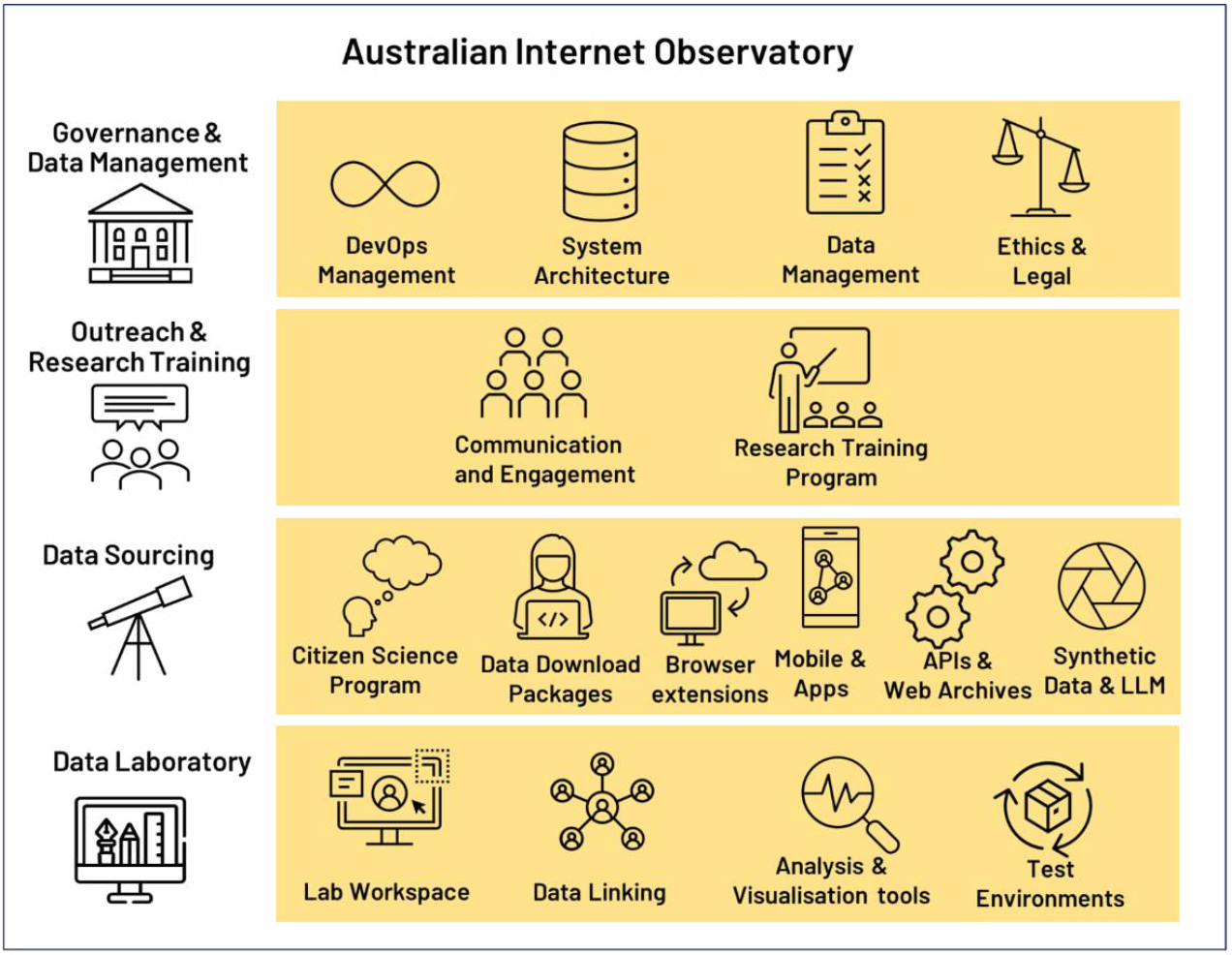

Figure 1. Planned Streams and Work Packages for the Australian Internet Observatory

Stream 1. Governance and Data management

Social data analysis is a relatively new field of data science which will require the application of existing best practice and new approaches to data management, ethics and research methods. The facility will be driven by best practice governance and ethics for data collection and management. 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. AIO 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 of social data is complex, involving issues of privacy, security, intellectual property, commercial control, bias in data and algorithms, ethics, noise and inaccuracies. AIO 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.

WP1. Development and Operations (DevOps) Management

Develop DevOps methodology and collaborative work spaces review and implementation for ongoing and integrated development and operations across the whole project. This will include setting up and maintaining: team workspaces and collaborative documentation; Infrastructure as Code (IaC); monitoring for key performance metrics; continuous integration and continuous delivery/deployment (CI/CD) systems; and project planning for next stage.

WP2. System architecture

Design and documentation of system architecture necessary to facilitate data collection and laboratory across the life of the project. This will include specification of a methodology that resourcefully integrates varying datasets, coupled with the specification of protocols that determine how internal/external parties interact with the datasets. The system architecture will also plan and implement HPC requirements, storage and authorisation and access systems.

WP3. Data management

Data management is a critical component of the project and will include setting up a data register and data work space protocols for heterogenous data and undertaking Indigenous data governance consultation and developing a framework in consultation with the HASS&I RDC. It will also include developing protocols for data stewardship such as access control, data sharing, and data lifecycle management. Documentation of guidelines for identifying responsible individuals or organisations for managing specific datasets, including protocols for data stewardship such as access control, data sharing, and data lifecycle management.

WP4. Ethics and legal frameworks

This WP involves a review and consultation on key ethical and legal issues, such as privacy and data use, and then developing and implementing an ethics framework within the platform and tools. This includes creating proforma ethics information, online guides, system changes based on the framework, and setting up a continuous improvement framework for ongoing review of the ethics implementation.

Stream 2. Community outreach and Research Training

WP5. Project communications and stakeholder engagement

This WP involves developing a communication and stakeholder engagement strategy, establishing a project website and social media accounts, and preparing for participation in events such as International Data Week in Brisbane, 2025 and a second co-design consultation. Following these activities, the project plan will be reviewed and revised as necessary based on the outcomes of the co-design consultation process.

WP6. Research training program

The program will provide training, tools and guides for Australian researchers across HASS and STEM disciplines, as well as for researchers in government, industry and civil society organisations. Following a sector-wide needs audit to identify challenges in digital dataset research, the training program will develop an advocacy strategy and skills uplift program for promoting research with social data via in person and online workshops and a resource hub for researchers, HDRs, and industry on principles and opportunities for integrating data science in HASS&I research.

Stream 3. Data sourcing

AIO will enable researchers access and analysis capability covering a range of mediated data. The data will primarily be sourced (1) from platforms, either through official or unofficial Application Programming Interfaces (APIs), (2) from users in the form of crowdsourced or donated data, or (3) brokered through connections with ADOReD/ADO, etc. The format of the assets in these datasets will be images, videos, audio files, or text. The structure of the datasets will vary and they will be associated with (1) static metadata characterising separate assets, (2) dynamic metadata representing engagement with the assets over time and space, (3) global identifiers facilitating the linking of assets across datasets. In addition to generation of these datasets from real-world sources, AIO will have the capability to generate synthetic data using generative AI and other relevant mechanisms.

WP7. Citizen Science program

The data donation model generally enrols participants to contribute to projects as ‘citizen scientists’ (Araujo et al., 2021). This way is similar to scientific projects that encourage patients to donate their health data, amateur meteorologists, naturalists or astronomers to contribute their observations to professional scholarly efforts. The citizen science program will encourage diverse participants and cover all types of participation and data donation methods at any stage of the process. This work package involves the creation of rigorous participant consent information and protocols, as well as the compilation of a register of ethics frameworks and protocols for data collection and management. It will also include a user needs analysis consultation with citizen scientists for data sourcing, the development of participant personas, story and workflow mapping for the data sourcing stream, and the creation of requirements for a Data Donation Registration System - a website and database for participants to register and join data donation projects.

WP8. Data sourcing through data download packages (DDP)

This WP builds infrastructure that exploits the requirement by the EU regulation GDPR, that digital media platforms operating in the EU (i.e. most large international platforms) must provide a facility for its users to download their personal data stored by the platform. This allows for a new approach to data collection when users of Netflix, TikTok, YouTube, etc, are supported by researchers to request their personal data from the platforms and, if they choose to do so, donate the data to for research purposes. A number of research initiatives,

primarily in Europe, have developed dashboards and tools to allow for this kind of citizen science inspired research. This WP involves a number of deliverables, ultimately building a system with a user-friendly dashboard where citizens can (1) access their personal data from the platforms they use in their daily lives; (2) explore that data to increase their data literacy and understand how their data is being used by multinational media platforms; and (3) donate anonymised versions of their personal data to research groups who are advertising their projects on the dashboard.

This work package will involve conducting an environmental scan of world-leading research initiatives, data sources, and tools for potential re-use, followed by a report and recommendations. It will also include the launch of a pilot focused on a single platform and a small group of citizen scientists, the design and decision of system architecture, and the expansion of the system to enable data collection from multiple platforms.

WP9. Data sourcing through browser extensions and plugins

Browser extensions can facilitate data acquisition from online platforms, by collecting 'in-browser' data from participating volunteers. The collected data are termed 'data donations', in this context meaning data gathered from (but not limited to) social media platforms, web applications, credit scoring services, and shopping online. Browser extensions can also gather data through self-facilitated background activities, such as querying online search engines. Browser extensions can be publicly downloaded from online extension store and the installation procedure may collect voluntarily supplied demographic information from participants. These optional responses are valuable in investigating whether different demographic groups encounter differing data but do not allow us to identify or track individual users, either on social media platforms, or elsewhere online. The browser extensions will be sensibly designed to not access any personal information from the user’s computer or online profile.

Both the data donations and the demographic information produced by this work package contribute directly to the creation of valuable social data, while the source code of the browser extensions constitute software artefacts that when made publicly available will assist similar projects in improving the observability of social media data.

WP10. Sourcing data from mobile devices

Most interactions with digital platforms are via mobile apps and interfaces so we need to be able to study them at this interface. Mobile software applications can facilitate data acquisition from social media platforms, by collecting 'screen-recorded' data from participating volunteers. The collected data are termed 'data donations', in this context relating to screen-recorded visual data automatically gathered directly from active social media platforms, as accessed on volunteer mobile devices. Exhibited with information pertaining to the research project, mobile software applications can be publicly downloaded from mobile app stores, as managed by the relative device's operating systems.

The installation procedure may collect voluntarily supplied demographic information from participants. These optional responses are valuable in investigating whether different demographic groups encounter differing data but do not allow us to identify or track individual users on social media platforms. The mobile software applications will be sensibly designed not to access any personal information from a given participant's mobile device. Both the data donations and the demographic information produced by this work package contribute directly to the creation of valuable social data, while the source code of the mobile software application will constitute a software artefact, that when made publicly available will assist similar projects in improving the observability of social media data.

This work package will involve scoping the necessary platform data acquisition requirements for selected social media platform apps and developing autonomous 'screen-recording' mechanisms for data acquisition. It will also include designing a mobile software application user interface, rigorously prototyping the application in various configurations, and releasing it to designated mobile app stores. Ongoing testing and maintenance will be

performed to anticipate changes to platforms that may affect the application's functionality. Additionally, a mobile version of the web archiving toolkit will be developed.

WP11. Data sourcing via platform APIs and Web Archiving

This work package builds on the previous investments by ARDC and university partners into systems such as the Digital Observatory, ADOReD and RAPID and the data produced. Although APIs are now more limited in what they can access and collect they still provide access to some platforms and collections. This WP will develop enhanced tools for collecting and analysing platform data via APIs, develop a Reddit databank and enhance the RAPID system to integrate new platform data. It will also involve development of a production-grade version of the NewsTalk prototype and NewsTalk databank. This WP will also repurpose existing web archiving tools and build an enhanced web archiving toolkit to support HASS researchers doing data collection for small projects. This will require the development of a set of web archive tooling designed to support the systematic use of web archives as a citable, reproducible method of web observations, both through manual and automatic browsing. It will address a number of problems ranging from usability of web archives for researchers, and the protocol and infrastructure necessary to cite web archives in academic work.

WP12. Creating synthetic data with LLMs

Synthetic data created via large language models is a relatively new 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). This can help to ensure transparency and understandability of algorithms and the use of AI for a first-pass understanding of data. This work package will involve investigating, identifying, and documenting synthetic information needs for large-scale evaluation of information access systems. It will also establish an open-source Large Language Model (LLM) for natural language querying/processing, integrate generative AI models that combine text and another modality, and implement tools for controlling a population of agents based on these models. Users will be able to control aspects of the agent personas, such as language style and stance on topics. Finally, it will integrate multi-modal generative AI models that combine text and several other modalities, with control of the subpopulations of agent personas.

Stream 4. Data Laboratory

AIO will provide an integrated laboratory of machine learning tools, visualisations and applications for researchers to work with and apply across diverse social data sources. These will be supported through enabling systems such as cloud computing and secure access. AIO will support the collection input of ‘messy’ data to be converted into more robust, archival-quality, open data standards. Data and analysis will be further enhanced and integrated with other data and tools through linking and standardised labelling.

WP13. Lab Workspace and Integrated Data

This work package will conduct a review and user needs analysis to determine researcher data, tool and access need, then will develop an online workspace to provide access to data and tools. The WP also includes harmonising the various types of data sourced via the other work packages based on linked data principles and agreed standards established across the HASS&I focus areas and identifying relevant vocabularies that could be applied to the social data collected to enable classification and interoperability.

WP14. Analysis and visualisation tools

Building on open source tools already developed by Australian researchers such as the Image Machine and Australian Text Analytics Platform (ATAP) at UQ and the QUT Digital Observatory open source ecosystem, as well as other open source tools, this WP will identify and provide access to key open source data analysis and

visualisation tools to support data cleaning, analysis and visualisation. The Workspace will also include scripts and packages to operate within emerging 'clean-room' environments that are increasingly provided by major platforms, many through the Inter-university Consortium for Political and Social Research (ICPSR). This WP will also 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 and visualisation tools such as graph, time and geospatial will also be provided. This work package will involve auditing types of image and data collections and their analysis needs, prioritising and planning the sequence of machine vision tool development in consultation with the researcher community, and integrating with the Lab Workspace to support secure dataset upload for processing and visualisation. Additionally, tools will be developed for clustering images according to emergent patterns, auditing and exploring how models classify, label and cluster images, and using image collections to produce synthetic text and images.

WP15. Test environments

This work package will identify possible application scenarios for generative AI, where HASS researchers require tools for generating and evaluating synthetic data. We will also identify scenarios where security threats might be an issue due to the misuse of generative AI. The Test Environment will provide a sandbox in which the impact of these misuse scenarios can be tested, as well as the effectiveness of proposed defences.

1.2. Budget

[redacted for publication]

1.3. Payment Schedule

[redacted for publication]

1.4. Project Partners & Subcontractors

Projects partners are: RMIT University, Queensland University of Technology, University of Queensland, University of Melbourne, Swinburne University and Deakin University. Collaborating organisations include research centres such as the ARC Centre of Excellence for Automated Decision-Making + Society, and the ARC Centre of Excellence for the Digital Child, and NCRIS facilities including the Australian Urban Research Infrastructure Network. International collaborators include the US National Digital Observatory and Smart Data Research UK.

ORGANISATION	SUBCONTRACTOR (SELECT BOX IF YES)	SUMMARY
RMIT University (lead)	☐	RMIT is a global university of technology, design and enterprise. It is a world leader in Art and Design; Architecture; Education; Engineering; Development; Computer Science and Information Systems; Business and Management; and Communication and Media Studies. AIO will be hosted within the College of Design and Social Context and the School of Media

		and Communication. In the 2021 QS Rankings by Subject RMIT Communications and Media Studies was ranked 3rd in Australia and remained in the top 50 globally, moving up three places. In the most recent Excellence in Research for Australia (ERA) ranking, RMIT was judged to be well above world standard (5) for the field of Communication and Media studies and above world standard (4) for the field of Cultural Studies and Design Practice and Management. RMIT hosts a number of major research centres and institutes including two ARC Centres of Excellence: The ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S), and the recently announced ARC Centre of Excellence Optical Microcombs for Breakthrough Science (Combs). ADM+S has been a driving force behind the development of AIO as part of the centres' aims to create the knowledge and strategies necessary for responsible, ethical, and inclusive automated decision-making. The ADM+S Centre is hosted at RMIT University with nodes located at eight other Australian universities including QUT, UQ, Melbourne and Swinburne, and partners around the world. It brings together universities, industry, government and the community to support the development of responsible, ethical and inclusive automated decision-making. RMIT is also partner on a number of research infrastructure projects including AURIN, and is a member of the ARDC. RMIT's enabling environment includes the RMIT AWS Cloud Supercomputing Hub (RACE Hub), a collaboration with Amazon Web Services (AWS) and AARNet to implement a dedicated commercial cloud supercomputing facility, enabling true scalable and elastic high-powered computing to support digital innovation. By combining the industry-leading cloud capabilities of AWS and the latest fibre network technologies from AARNet, RMIT is set to access tremendous connectivity and HPC processing power and provide seamless access to all our researchers, academics, students and industry partners.
Queensland University of Technology (QUT)	☑	QUT hosts a node of ADM+S and the Australian Research Council Centre of Excellence for the Digital Child and is a member of the ARDC. The Digital Media Research Centre (DMRC), led by Prof Dan Angus is a global leader in digital humanities and social science research, with a focus on communication, media and the law. It addresses local, national and global challenges at the forefront of digital transformation. The DMRC is a member of the global Network of Centers, a group of academic institutions that conduct interdisciplinary research on the development, social impact, policy implications and legal issues surrounding the internet. The AIO node at QUT will be co-located with the Australian Digital Observatory (ADO) and within the Digital Media Research Centre (DMRC), led by Prof Dan Angus.

		The QUT Digital Observatory is a QUT Research Infrastructure facility and an active contributor to the Australian Digital Observatory (ADO) project co-funded by ARDC.
University of Queensland (UQ)	☑	UQ is a member node of both the ARC Centre of Excellence for Automated Decision-Making and Society and the ARC Centre of Excellence for the Digital Child. It is also a partner on HASS RDC projects LDaCA and the IRDC. The AIO node will be based in the Digital Cultures and Societies Research Centre at UQ led by Prof Nic Carah. The DCSRC brings together researchers to pose urgent questions about how we build inclusive, flourishing and sustainable societies and cultures using digital technologies and media. UQ is a member of the ARDC and a host university for ARDC staff.
The University of Melbourne	☑	The University of Melbourne is a partner on a number of NCRIS facilities including AURIN and the Australian Digital Observatory and is a node in the ARC Centre of Excellence for Automated Decision-Making + Society. AIO will be working with the School of Computing and Information at the University led by Prof Chris Leckie, Director of the Academic Centre of Cyber Security Excellence (ACCSE). The School of Computing and Information Systems is an international research leader in computer science, information systems and software engineering. We are focused on delivering impact in the following key areas: ● Artificial intelligence: deep learning, data mining, machine learning, natural language processing, and agent-based systems. ● Computer science: programming languages, algorithms, distributed computing, and cybersecurity and cryptography. ● Human-Computer Interaction: issues arising from humans interacting with emerging technologies. ● Information systems: The study of the interplay between information technology, its users (people), and the operations (business processes) through which this technology is used.
Swinburne University	☑	Swinburne University is a node of ADM+S and also a partner on the IRISS/CADRE project and the NCRIS facility Astronomy Australia. The AIO node led by Prof Anthony McCosker, will be based in the Social Innovation Research Institute (SIRI) within the School of Social Sciences, Media, Film and Education. It is a leading research centre on social innovation in the digital economy. It creates solutions to complex social problems using co-design, new technologies and data analytics techniques. SIRI researchers work at the intersection between social challenges and the potential of technology, striving for greater social equity. SIRI includes major research infrastructure projects Analysis & Policy Observatory, the Social Data Analytics Lab (SoDA Lab). Swinburne is a member of ARDC.

Deakin University	☑	Deakin University is a node in the ARC Centre of Excellence for the Digital Child. In addition, the Faculty of Arts and Education where the AIO node will be based is host to theResearch for Educational Impact (REDI) and the Alfred Deakin Institute for Citizenship and Globalisation. REDI is a leading national and international centre in educational research. REDI takes an outward-facing approach to delivering and translating high quality research and research training into outcomes that are relevant and meaningful to communities. Its research is focussed around four distinctive themes, and is led by renowned scholars in collaboration with highly active and successful educational researchers from a number of disciplines, as well as from the Centre for Research in Assessment and Digital Learning (CRADLE). The Alfred Deakin Institute for Citizenship and Globalisation (ADI) which is home to many of the Faculty's social science and humanities researchers, working on issues associated with racism, social inequalities, counter terrorism and youth studies. The AIO node will be led by Associate Professor Matteo Vergani, a social scientist in ADI working at the intersection between online and real world hateful behaviours. He will be supported by Professor Andrea Witcomb, Associate Dean Research in the Faculty and Professor Julian Sefton-Green, leader of the Deakin node of the ARC Centre of Excellence for the Digital Child.

1.5. Project team roles and responsibilities

[redacted for publication]

1.6. Governance

Project planning, monitoring and reporting

The Project Manager will be responsible for project management planning, monitoring and reporting as well as scheduling, budget and coordination of the project team and relationships.

Stakeholder management

The Australian Social Data Observatory will be managed as a national facility, with a management structure reflecting a broad range of stakeholders including the HASS&I RDC, ARDC, other NCRIS facilities, university partners, Centres of Excellence, and industry partners.

The AIO Project Team operations will be based on the model of Distributed Research Infrastructure (DRI), a 'national association of geographically-separated distinct entities that jointly perform, facilitate or sponsor basic or applied scientific research'[1] (OECD 2014). AIO will operate as one distributed node per partner-RMIT, QUT, UQ, Melbourne, Swinburne and Deakin with a coordinating role played by the lead organisation (RMIT) where the PM

[1] OECD 2014, International distributed research infrastructure, https://www.oecd.org/science/international-distributed-research-infrastructures.pdf

will be based. Stakeholder management will include: managing partner relationships; coordinating and responding to the project steering group and community advisory group; coordinating the project team. A technical advisory group will be established if required.

FAIR and CARE principles - governance arrangements

As part of the Governance and Data Management stream of work, AIO 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. Work will include consultation with Indigenous stakeholders and the development of a governance framework for Indigenous data within AIO.

Steering committee

The AIO Steering Committee will oversee the implementation and performance of the project. The steering committee will be drawn from partner leaders, the ARDC and other key stakeholders.

The committee will be chaired by the academic lead and provide approval of the key deliverables of the project – primarily the overall project deliverables outlined in work package 1, including project plan, periodic reporting and financial reports.

The committee will meet quarterly, with secretariat support provided by the project management group as part of Work Package 1. The project lead and chair of the AIO steering committee will also participate in joint HASS-I program governance as required.

The proposed membership of the committee will include:

- Partner leads (RMIT, QUT, UQ, Melbourne, Deakin, Swinburne)
- HASS-I RDC Program Manager (ARDC)
- Other key stakeholders as required

AIO Steering Group members (TBC)

Julian Thomas, Program lead (Chair)

Daniel Angus, QUT Lead

Nicholas Carah, UQ Lead

Chris Leckie, UoM Lead

Anthony McCosker, Swinburne Lead

Matteo Vergani, Deakin Lead

Jenny Fewster, HASS&I Program Director, ARDC

Program Manager, Ex officio

Other key stakeholder representatives (TBC)

1.7. Implementation & Timing - Milestones and Deliverables

The following are the agreed milestones and deliverables for the project.

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
Stream 1. Governance and Data management				
WP1 Development and Operations (DevOps) Management				
1.1. DevOps methodology review and selection of collaborative work spaces for ongoing and integrated development and operations across the whole project.	RMIT		Oct 2024	Mar 2025
1.2. Basic CI/CD pipelines set-up and tested for critical functionalities.	RMIT		Jan 2025	Jun 2025
1.3. Onboarding and initial configurations collaboration and documentation material available for onboarding and initial configurations.	RMIT		Jan 2025	Jun 2025
1.4. Steer Co sign off on WP 1.1, 1.2, 1.3	*SC*		*Jun 2025*	*Jun 2025*
1.5. Introduce basic Infrastructure as Code (IaC) for provisioning essential resources	RMIT		Jul 2025	Mar 2026
1.6. Set up initial monitoring for key performance metrics.	RMIT		Jul 2025	Jun 2026
1.7. Steer Co Sign Off on WP 1.5 & 1.6	*SC*		*Jun 2026*	*Jun 2026*
1.8. Implement DevOps and Evaluation WPs based on revised project plan (WP 5.5)	RMIT		Oct 2026	Jun 2028
WP2. System architecture				
2.1. Draft design and documentation of system architecture necessary to facilitate data collection and laboratory.	QUT		Oct 2024	Mar 2025
2.2. Review and document HPC requirements and options	QUT	Consultation with ARDC	Oct 2024	Mar 2025

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
2.3. Specification report on a methodology that integrates varying datasets (such as those compiled through data acquisition phases, or through external parties), coupled with the specification of protocols that determine how internal/external parties interact with the datasets.	QUT		Oct 2024	Jun 2025
2.4. Review authorisation and access options in consultation with ARDC and provide recommendations. (Informed by WP3.2)	RMIT	Consult ARDC	Oct 2024	Jun 2025
2.5. Steer Co sign off on WP 2.1, 2.2, 2.3, 2.4	*SC*		*Jun 2025*	*Jun 2025*
2.6. Storage system or repository set up (depending on review of data governance and storage design)	RMIT	Consult ARDC, use storage systems TBC	Jul 2025	Mar 2026
2.7. Secure access system with required levels and permissions set up	QUT	Consult ARDC, re access systems	Jul 2025	Mar 2026
2.8. Steer Co sign off on WP 2.6, 2.7	*SC*		*Jun 2026*	*Jun 2026*
2.9. Revise and implement system architecture based on Revised Project Plan (WP 5.5)	QUT		Jul 2026	Jul 2028
WP3. Data management				
3.1. Data register and data work space protocols completed: documentation of the types of data to be collected, their sources, storage and access requirements, as well as data distribution, backup and security plans.	Swin	Consult ARDC	Oct 2024	Jun 2025
3.2. Agreed principles and project plan for HASS&I RDC Integration planning	RMIT	Consult ARDC	Oct 2024	Jun 2025
3.3. Steer Co sign off on WP 3.1, 3.2	*SC*		*Jun 2025*	*Jun 2025*
3.4. Draft Indigenous data governance consultation and governance framework completed: documentation of consultation process, guidelines for data handling and protocols for	Swin	Consult ARDC	Oct 2024	Jun 2026

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
Indigenous Data in collaboration with the Indigenous Data Network.				
3.5. Data stewardship register and protocols documentation completed	Swin		Jul 2025	Jun 2026
3.6. Steer Co sign off on WP 3.4, 3.5	*SC*		*Jun 2026*	*Jun 2026*
3.7. Implement revised data management WPs based on revised Project Plan WP 5.5	Swin		Oct 2026	Jun 2028
WP4. Ethics and legal frameworks				
4.1. Consultation and Review of ethics and legal framework key issues ie privacy, sensitive data, reuse of commercial data etc.	RMIT	Consult ARDC	Oct 2024	Jun 2025
4.2. Develop proforma ethics information for research projects and tools	RMIT	Consult ARDC	Jan 2025	Jun 2025
4.3. Steer Co sign off on WP 4.1, 4.2	*SC*		*Jun 2025*	*Jun 2025*
4.4. Implementation plan for ethics and legal framework within platform and tools	RMIT		Jul 2025	Jun 2026
4.5. SC sign off on WP 4.4	*SC*		*Jun 2026*	*Jun 2026*
4.6. Review of ethics framework implementation and continuous improvement framework set up.	RMIT		Oct 2026	Jun 2028
Stream 2. Community outreach and Research Training				
WP5. Project communications				
5.1. Develop communications and stakeholder management plan	RMIT	Consult ARDC	Oct 2024	Mar 2025
5.2. Implement comms and engagement activities	RMIT	Consult ARDC	Oct 2024	Jun 2026
5.3. Steer Co Sign off on WP 5.1, 5.2	*SC*	*Consult ARDC*	*Jun 2025*	*Jun 2025*

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
5.4. Prepare for and participate in Co-design consultation	RMIT	ARDC Engagement Team	Jan 2026	Jun 2026
5.5. Revise Project Plan based on co-design consultation (WP5.4)	RMIT	ARDC PMO Team	Jul 2026	Sep 2026
5.6. Steer Co sign off on WP 5.4, 5.5	*SC*	*PM*	*Sep 2026*	*Sep 2026*
5.7. Implement revised Comms and engagement strategy based on WP5.5	RMIT	Consult ARDC	Oct 2026	Jun 2028
WP6. Research training program				
6.1. Conduct a sector-wide needs audit to identify challenges in digital data research, foster industry partnerships, and bridge gaps between data and social science, enhancing tool engagement and understanding industry needs.	Deakin	ARDC Research Training expertise	Oct 2024	Jun 2025
6.2. Research training plan developed	Deakin	Consult ARDC	Jan 2025	Jun 2025
6.3. Steer Co Sign off on WP 6.1 and 6.2 progress to date	*SC*	*Consult ARDC*	*Jun 2025*	*Jun 2025*
6.4. Prepare and deliver program of training and workshops on principles, opportunities, tools and methods for integrating data science in HASS&I research	Deakin	ARDC Training team support	Jan 2025	Jun 2026
6.5. Develop online information and training resources on ethics, data acquisition, analysis, manipulation, linking and sharing	Deakin	ARDC Training team support	Jan 2025	Jun 2026
6.6. Describe and document exemplar research data workflows that select and use suitable data tools as a guide to researchers in their choice of tools for their research	Deakin		Jan 2026	Jun 2026
6.7. Steer Co sign off on WP 6.4, 6.5, 6.6	*SC*		*Jun 2026*	*Jun 2026*
6.8. Implement research training program based revised project plan WP5.6	Deakin	ARDC Training team support	Oct 2026	Jun 2028

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
Stream 3. Data sourcing				
WP7. Citizen Science program				
7.1. Citizen science principles and ethics frameworks and protocols review and develop best practice guidelines	UQ		Oct 2024	Mar 2025
7.2. Develop participant personas, story and workflow mapping for Data sourcing stream	UQ		Oct 2025	Mar 2025
7.3. Develop requirements for Data Donation Registration System (DDRS)	UQ		Mar 2025	Jun 2025
7.4. Steer co sign off on WP 7.1, 7.2, 7.3	*SC*		*Jun 2025*	*Jun 2025*
7.5. Develop Beta DDRC V1.0	UQ		Jul 2025	Sep 2025
7.6. Launch DDRS V2.0	UQ		Sep 2025	Dec 2025
7.7. Develop DDRS V3.0	UQ		Jan 2026	Jun 2026
7.8. Steering Co sign off on WP 7.5, 7.6, 7.7	*SC*		*Oct 2026*	*Oct 2026*
7.9. Continuous improvements to DDRS based on revised project plan WP 5.5	UQ		Oct 2026	Jun 2028
WP8. Data sourcing through data download packages (DDP)				
8.1. Complete report and recommendation of international DDP research initiatives, data sources and tools.	QUT		Oct 2024	Mar 2025
8.2. Develop DDP pilot system architecture and recruitment program	QUT		Apr 2025	Jun 2025
8.3. Steer co sign off WP 8.1, 8.2	*SC*		*Jun 2025*	*Jun 2025*
8.4. V1 launched, focused on a single platform and a small group of citizen scientists.	QUT		Aug 2025	Dec 2025

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
8.5. Expanded to v2 of the DDP infrastructure, enabling data collection from multiple platforms.	QUT		Jan 2026	Jun 2026
8.6. Steer Co Sign off on WP 8.4, 8.5	*QUT*		*Jun 2026*	*Jun 2026*
8.7. Expand the system with capability for research group interface and communication (V3) based on Revised project plan WP 5.5	QUT		Oct 2026	Jul 2028
WP9. Data sourcing through browser extensions and plugins				
9.1. Scope necessary platform data acquisition requirements of selected social media platforms	QUT		Oct 2025	Mar 2025
9.2. Design browser-based extension user interface and compile participant forms that facilitate the collection of demographic data.	QUT		Jan 2025	June 2025
9.3. Rigorously prototype the designed browser extension within a variety of configurations, ensuring its fidelity across the full specification of its functionality.	QUT		Jan 2025	June 2025
9.4. Steer Co Sign off on WP 9.1, 9.2, 9.3	*SC*		*Jun 2025*	*Jun 2025*
9.5. Release the browser extension to a designation of web browsers,	QUT		Jul 2025	Sep 2025
9.6. Anticipate changes to platforms that may affect the browser extension's functionality, through the performance of ongoing testing and maintenance.	QUT		Oct 2025	Jun 2026
9.7. Steer Co sign off on WP 9.5, 9.6	*SC*		*Jun 2026*	*Jun 2026*
9.8. Revise and expand browser extensions based on review (WP5.5)	QUT		Oct 2026	Jun 2028
WP10. Data sourcing from mobile devices				

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
0.1. Scope necessary platform data acquisition requirements for selected social media platform apps, and develop autonomous 'screen-recording' mechanisms that facilitate acquisition..	QUT		Jan 2025	June 2025
0.2. Design mobile software application user interface, including compilation of participant forms that facilitate the collection of demographic data.	QUT		Oct 2024	Mar 2025
0.3. Rigorously prototype the designed mobile software application within a variety of configurations, ensuring its fidelity across the full specification of its functionality.	QUT		Jan 2025	June 2025
0.4. Steer Co sign off WP 10.1, 10.2, 10.3	*SC*		*Jun 2025*	*Jun 2025*
0.5. Release the mobile software application to a designation of mobile app stores and anticipate changes to platforms X and Y that may affect the application's functionality, through the performance of ongoing testing and maintenance	QUT		Jul 2025	Jun 2026
0.6. Steer Co Sign off on WP 10.5	*SC*		*Jun 2026*	*Jun 2026*
0.7. Revise and expand mobile data donation systems based on revised project plan (WP5.5)	QUT		Jul 2026	Jun 2028
0.8. Mobile Web Archive Prototype: Development of a mobile version of the web archiving toolkit.	QUT		Jul 2026	Jun 2028
WP11. Data sourcing via platform APIs and access to existing databanks.				
1.1. Scope requirements and availability of platform APIs and Web archiving	QUT		Oct 2024	Dec 2025
1.2. Develop enhanced tools for collecting and analysing platform data via APIs	QUT		Jan 2025	Jun 2025

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
1.3. Collection Management Tools and Web Archive File Index	QUT		Jan 2024	Jun 2025
1.4. Integrate existing RAPID platform for collecting data from online data sources through their APIs in real-time	Melb		Oct 2024	Jun 2025
1.5. Steer Co Sign off on WP 11.1, 11.2, 11.3, 11,4	*SC*		*Jun 2025*	*Jun 2025*
1.6. Production-grade version of the NewsTalk prototype developed	QUT		Jan 2025	Jun 2026
1.7. Enhance RAPID to develop interfaces to collect data from Mastodon instances GDELT for news media (free data sources) or other similar data sources.	Melb		Jul 2025	Jun 2026
1.8. Development of a Web Archive Server	QUT		Jul 2025	Jun 2026
1.9. Steer Co Sign off on WP 11.6, 11.7, 11.8	*SC*		*Jun 2026*	*Jun 2026*
.10. Implement new WPs based on revised project plan WP5.5	QUT		Oct 2026	Jun 2028
.11. Develop Web Archive Command Line Data Extraction tool	QUT		Oct 2026	Jun 2028
WP12. Create synthetic data with LLMs				
2.1. Establish open source LLM for natural language querying/processing, for example LLAMA2 or similar. Provide examples of how to use for potential use-cases identified in WP15.1	Melb		Oct 2024	Mar 2025
2.2. Integrate generative AI models that combine text and one other modality (e.g., speech, images or video).	Melb		Jan 2025	Jun 2025
2.3. Steer Co sign off on WP 12.1, 12.2	*SC*		*Jun 2025*	*Jun 2025*
2.4. Implement tools for controlling a population of agents based on generative AI models, where users can control aspects of the personas of the	Melb		Jul 2025	Jun 2026

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
agent population, e.g., style of language, stance on topics, response to interactions.				
2.5. Steer Co sign off on WP 12.4	*SC*		*Jun 2026*	*Jun 2026*
2.6. Integrate multi-modal generative AI models that combine text and several other modalities (e.g., some combination of speech, images or video) with control of the subpopulations of agent personas.	Melb		Jul 2026	Jun 2028
Stream 4. Data Laboratory				
WP13. Lab Workspace and Data Integration				
3.1. Review of user needs for data access, training and analysis	UQ		Oct 2024	Mar 2025
3.2. Develop requirements and specifications for Lab Workspace MVP	UQ		Jan 2025	Mar 2025
3.3. Set up Beta Lab Workspace	UQ		Apr 2025	Jun 2025
3.4. Develop data workflow plan between Data sourcing stream and Data Laboratory	QUT		Apr 2025	Jun 2025
3.5. Steer Co sign off on WP 13.1, 13.2, 13.3.4	*SC*		*Jun 2026*	*Jun 2026*
3.6. Lab Workspace V2.0 available for internal users	UQ		Jul 2025	Dec 2025
3.7. Identify and develop integration plan for relevant vocabularies and identifiers for linked data interoperability	Swin		Jul 2025	Dec 2025
3.8. Launch Lab Workspace V3.0 for external users	UQ		Jul 2025	Dec 2025
3.9. Integrate Key vocabularies and identifiers in Lab Workspace and Data Flows	UQ		Jan 2026	Jun 2026
10. Steer co sign off on WP 13.6, 13.7, 13.8,3.9	*SC*		*Jun 2026*	*Jun 2026*
11. Develop Lab Workspace V3.0 and linked data based on project plan WP 5.5	UQ		Oct 2026	Mar 2028

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
WP14. Analysis and visualisation tools				
4.1. Review and prioritisation of existing analysis and visualisation tools and gap analysis	QUT	Consult ARDC	Oct 2024	Mar 2025
4.2. Prioritise and plan sequence of machine vision tool development in consultation with researcher community.	UQ		Oct 2024	Mar 2025
4.3. Select ML image tools and open source tools to be made available via Lab Workspace V1.0 (WP 13.4)	UQ		Apr 2025	Jun 2025
4.4. Steer Co sign off on WP 14.1,14.2.14.3	*SC*		*Jun 2025*	*Jun 2025*
4.5. Integrate additional open source tools into Lab Workspace	UQ		Jul 2025	Jun 2026
4.6. Develop and implement tools for manual annotation and labelling of images to create culturally-specific 'ground truth' to inform automated labelling and/or clustering	UQ		Jul 2025	Jun 2026
4.7. Integrate graph visualisation tool into Lab Workspace	UQ		Jan 2026	Jun 2026
4.8. Steer co sign off on WP 14.5,14.6, 14.7	*SC*		*Jun 2026*	*Jun 2026*
4.9. Develop and Integrate adiditional tools based on revised project plan WP 5.5	UQ		Oct 2026	Jun 2028
10. Implement key tools for clustering images according to emergent patterns	UQ		Oct 2026	Dec 2026
11. Develop tools and interface for auditing and exploring how models classify, label and cluster images and produce synthetic text and images.	UQ		Oct 2026	Dec 2026
WP15. Test environments				
5.1. Investigate, identify and document appropriate application scenarios where the safety and security of generative AI models needs to be evaluated.	Melb		Oct 2024	Mar 2025
5.2. Implement a proof-of-concept test environment to demonstrate the effectiveness of generative AI tools in an illustrative application scenario.	Melb		Oct 2024	Jun 2025
5.3. Steer Co sign off on 15.1, 15.2	*SC*		*Jun 2025*	*Jun 2025*

DELIVERABLE / WORK PACKAGE	RESPONSIBILITY (Org)	ARDC RESOURCES REQUIRED	START DATE	FINISH DATE
5.4. Integrate tools for generative AI test range that implement a range of tests for known vulnerabilities in generative models, e.g., model jailbreaking attacks.	Melb		Jul 2025	Jun 2026
5.5. Steer Co sign off on 15.4	*SC*		*Jun 2026*	*Jun 2026*
5.6. Implement revised Test environment program based on project plan WP 5.5	Melb		Oct 2026	Jun 2028
5.7. Integrate tools for generative AI test range that implement different types of detection and mitigation techniques against misuse scenarios for generative AI models.	Melb		Oct 2026	Jun 2028
5.8. Integrate tools for generative AI test range that implement multi-modal attacks against generative AI models (e.g., attacks that exploit more than one type content such as text, images etc), and implement proposed detection and mitigation techniques from the research literature.	Melb		Oct 2026	Jun 2028

1.8. Visual Summary Overview

Please refer to Appendix B for a visual summary of the project.

1.9. Assumptions

The following assumptions are made in order to deliver successful project outcomes.

ITEM #	CATEGORY (Scope/cost/quality)	DESCRIPTION

1	Schedule	Recruitment will take 3 months to hire and onboard staff
2	Schedule	Subcontracts will be completed by Oct 2024 so work will begin after that.
3	Scope and quality	Citizen scientists will actively participate in data donation research projects
4	Scope and quality	Platforms will remain stable enough for data capture over a required period
5	Scope and Quality	Skilled human resources will be available to fulfil project roles and responsibilities.
6	Cost	The cost of labour will remain within the allocated budget for the project.

1.10. (Inter)dependencies

DEPENDENCY	RELATIONSHIP TO / IMPACT ON PROJECT	HOW AND WHO WILL MANAGE THE DEPENDENCY
ARDC cloud computer and storage resources	May impact data analysis	Technical Lead, QUT. A range of cloud compute and storage will be identified.
ARDC services	Will provide support and advice	PM, RMIT. The PM will work with ARDC staff to identify services available for use on the project.
Access and authorisation systems	This will be a critical technology but various robust systems exist	Data scientist, RMIT. There is a dedicated sub WP on Access.
Access to skilled staff to support the infrastructure	This will be critical.	The Project Manager at RMIT will manage this including sourcing contractors if required for short term periods while recruitment is happening.

Access to particular ARDC expertise	Will provide support and advice	PM, RMIT. The PM will work with ARDC staff to identify services available for use on the project.
Collaboration with other HASS&I RDC programs to identify common infrastructure needs	Will provide support and advice	PM, RMIT. The PM will work with other HASS RDC staff to identify services available for use on the project.

1.11. Risks

In the Controls/Mitigation Strategy section include what preventative actions you plan to take and/or actions you might take should the preventative actions fail to control the risk (i.e. what's your plan B?). These might include applying other in-kind resources, reviewing the plan and reducing scope etc.

Risk Rating Key

		Consequence				
		Insignificant (1)	**Minor (2)**	**Moderate (3)**	**Major (4)**	**Significant (5)**
Likelihood	**Almost certain (5)**	**5**	**10**	**15**	**20**	**25**
	Likely (4)	**4**	**8**	**12**	**16**	**20**
	Possible (3)	**3**	**6**	**9**	**12**	**15**
	Unlikely (2)	**2**	**4**	**6**	**8**	**10**
	Rare (1)	**1**	**2**	**3**	**4**	**5**

RISK	**IMPACT TYPE AND HOW WILL IMPACT PROJECT** (Scope/cost/quality/	**CONTROLS/MITIGATION STRATEGY**	**RESIDUAL RISK RATING** (after controls are in place)	**RISK OWNER**

	schedule)			
Difficulty recruiting specialised staff in software development and machine learning engineering roles.	Scope and schedule	The risk will be mitigated because the AIO infrastructure is led by researchers working in national research programs and centres that have developed a pipeline of talent and experience developing digital tools and research software. The distribution of responsibilities to different staff members, such as reflected in the distribution of leadership for the different WP leads in the project proposal will reduce the impacts from such risk, along with the nomination of a backup- team, and leave and contingency planning.		JT
Difficulty accessing digital data from platforms	Scope and schedule	This risk will be mitigated by several factors including: international legal reforms requiring platforms to provide data for user and public observability; the infrastructure emerging from a mature research field that has developed durable approaches to sourcing data from digital platforms using a variety of approaches. The most durable of these is citizen science and data donation approaches that are ethical, legal and generate a strong public mandate for observability.		DA
Rapid changes in the digital sector requiring revised project plans	Scope	We mitigate this risk by building on mature approaches in our research fields and ensuring the infrastructure is integrated with cutting-edge programs of research that enable forward-planning and anticipation of changes. Continuous development and integration methodologies designed to cope with this.		JT
Activity attracts legal attention from internet companies	Scope and Schedule	A strong governance and ethics framework		JT

Dependencies between WPs	Schedule	Continuous development and integration methodologies designed to cope with this.		DS and DA

**Only include risks that have a rating of greater than 14 to the project.*

1.12. Outputs and Outcomes Monitoring and Evaluation Plan

The indicators below specify what will be measured in the project M&E process in order to assess whether, and to what extent, the project's key intended outputs outcomes have been achieved.

End of Project Outputs are the deliverables achieved as part of the project.

End of Project Outcomes are the direct changes that occur from the outputs of the project that can be achieved within the timeframe of the investment.

OUTPUT	INDICATOR/S	MEASURE	DATA SOURCE/S	TIMELINE FOR DATA COLLECTION	RESPONSIBILITY	BASELINE
[Output delivered as part of the project]	**[An indicator that will demonstrate the success of the outcome]**	**[A measure of the indicator, that shows the change]**	**[Where will you source the data to support the indicator? How will you collect the indicator data? Is it quantitative / qualitative e.g. surveys, interviews]**	**[when will the data be collected and how often]**	**[Who will be collecting this information]**	**[Where available, what is the baseline Indicator for the outcome]**
e.g. Example: Delivery of research platform	Demand for use, number of citations of the platform	% subscribed/oversubscribed, # of citations by publications by authors using the platform	Platform data, Data from DOI cites.	6 monthly	Project manager	No baseline
EOP Output 1. Data sourcing tools	Researcher use of data sourcing tools	# of registered participants; # of extension downloads; quantity of	Site metrics;; tool download figures; data collection system; # researchers	Annual monthly	Program manager and Technical lead	No baseline

OUTPUT	INDICATOR/S	MEASURE	DATA SOURCE/S	TIMELINE FOR DATA COLLECTION	RESPONSIBILITY	BASELINE
		data collected for different tools; Evidence that data sourcing tools directly contributing to research projects	as registered users; use cases demonstrating value			
EOP Output 2. Data laboratory tools	Researcher use of data sourcing tools	# Registered researchers using tools and Lab Workspace; Evidence that data lab is directly contributing to research projects; # Workshop participants	Site metrics; Use cases demonstrating value	Annual	Program manager	No baseline
EOP Output 3. Capacity building training and resources	Researchers completed training or using resources	# of workshops offered; # participants completed workshops; # page views for online content	Workshop registrations; Workshop evaluations; Site metrics	Annual	Program Manager	No baseline

OUTCOME	INDICATOR/S	MEASURE	DATA SOURCE/S	TIMELINE FOR DATA COLLECTION	RESPONSIBILITY	BASELINE
[The change you intend to achieve as a result of the outputs of the project (please indicate)]	**[An indicator that will demonstrate the success of the outcome]**	**[A measure of the indicator, that shows the change]**	**[Where will you source the data to support the indicator? How will you collect the indicator data? Is it quantitative / qualitative e.g. surveys, interviews]**	**[when will the data be collected and how often]**	**[Who will be collecting this information]**	**[Where available, what is the baseline Indicator for the outcome]**
EOP Outcome 1. Adoption of infrastructure tools and/or data	Researchers and projects planning to or currently using tools	# of researchers and projects using tools or planning to use tools	# Registered users; # workshop participants; survey of researchers	Annual	PM	No baseline
EOP Outcome 2. Skills uplift for researchers	Training and engagement by researchers	# participants in workshops # users of online resources	Workshop registrations Google or Site analytics	Annual	PM	No baseline
EOP Outcome 3. Making it easier for researchers to source mediated data	User perceptions of increased availability of mediated data (via project)	Users report that mediated data has become more available in survey responses	Survey of community/user base	End of project	PM and Research leads	No baseline

OUTCOME	INDICATOR/S	MEASURE	DATA SOURCE/S	TIMELINE FOR DATA COLLECTION	RESPONSIBILITY	BASELINE
		and/or case studies				

1.13. Communications & engagement

AUDIENCE GROUP	WHY IS THIS AUDIENCE IMPORTANT FOR THE PROJECT?
University Research Centres and Institutes and research organisations in other sectors including: ARC Centre of Excellence for Automated Decision-Making + Society (ADM+S), ARC Centre of Excellence for the Digital Child, Digital Ethnography Research Centre (RMIT), Digital Cultures & Societies (UQ), Digital Media Research Centre (QUT), Social Innovation Research Institute (Swinburne) and the Centre for Research in Assessment and Digital Learning (Deakin), Alfred Deakin Institute for Citizenship and Globalisation (Deakin), Centre for Resilient and Inclusive Societies (Deakin), Academic Centre of Cyber Security Excellence (UoM), Centre for Artificial Intelligence and Digital Ethics (UoM), Cybernetics Institute (ANU).	Researchers at research centres and institutes will be key end users of the Observatory, will be kept informed throughout and may also contribute to building the project.
Media organisations	Media organisations will be interested in reporting on research findings and some may also be key users of the Observatory.
Industry including corporates in the technology and internet sectors	Industry groups will be interested in the research findings and some may also be users of the Observatory. Some may be directly impacted by the findings.
Civil society organisations concerned with online safety including: ACCAN, Humanitech, Foundation for Alcohol Research and Education (FARE), community organisations collecting online hate data (e.g., Islamophobia Register Australia, Executive Council of Australian Jewry, Asian Australian Alliance, Call it out)	Civil society organisations will be end users of the Observatory and will be kept informed throughout the project.
Federal and State Government Departments and Agencies:	Government departments and agencies will be key end users of the platform and pathway to impact
GLAM Institutions and Research Infrastructure Facilities and user groups including: National Library of Australia (NLA), NSLA, AURIN, ARDC HASS&I RDC and People and Planet thematics	Other research infrastructure facilities and their user groups will contribute to building the project, may be end users of the

	Observatory and will be kept informed throughout.

2. GLOSSARY OF TERMS

TERM	DESCRIPTION
API	Application Programming Interface
CS	Citizen Science
DD	Data Donation
Digital trace data	Digital traces left by individuals when they act or interact online
DDP	Data Download Packages
DevOps	Development and Operations: software development approach combining agile and waterfall methods into a continuous integration and development cycle.
LLM	Large Language Models
ML	Machine Learning

3. CHANGE CONTROL (for ARDC information only)

Approval of this Project Plan will comprise the baseline for the project. Changes to any of the following are considered variances:

- Project details
- Project outcomes & aims
- Budget
- Project partners
- Project team roles and responsibilities
- Governance
- Milestones and deliverables
- (Inter)dependencies

Variances to the Project Plan require endorsement by the Steering Committee and then ARDC approval. If approved, the Project Plan will be revised and project reports from that point forward will report project progress against the revised Project Plan, not the original.
To request a variance:

1. The Steering Committee submits a request to ARDC for variance to the approved project plan.
2. ARDC reviews the changes and advises the project manager or project lead of the outcome.

4. APPENDICES

[redacted]