Media and Communication
2026 • Volume 14 • Article 12090
https://doi.org/10.17645/mac.12090

ARTICLE

Open Access Journal

Building National Research Infrastructure for Digital Platform Observability at Scale

Bogdan Mamaev [1], Daniel Angus [2], Amanda Lawrence [3], and Richard Sinnott [4]

[1] School of Humanities and Social Sciences, Deakin University, Australia
[2] Digital Media Research Centre, Queensland University of Technology, Australia
[3] School of Media and Communication, RMIT University, Australia
[4] School of Computing and Information Systems, University of Melbourne, Australia

Correspondence: Richard Sinnott (rsinnott@unimelb.edu.au)

Submitted: 2 February 2026 **Accepted:** 22 June 2026 **Published:** 29 July 2026

Issue: This article is part of the issue "Open Research Infrastructures and Resources for Communication and Media Studies" edited by Silke Fürst (University of Zurich), Johannes Breuer (Center for Advanced Internet Studies / University of Duisburg-Essen), Erik Koenen (University of Bremen), Dimitri Prandner (Johannes Kepler University of Linz), Christian Schwarzenegger (University of Bremen), and Christian Strippel (Weizenbaum Institute), fully open access at https://doi.org/10.17645/mac.i504

Abstract

Digital platforms have fundamentally transformed public communication, yet the capacity to study them is increasingly threatened by restricted data access, opaque algorithmic architectures, and complex legal barriers. Addressing these challenges requires a paradigm shift from isolated, project-based solutions toward shared, national-scale research infrastructures. We present the Australian Internet Observatory (AIO) as a case study of a multi-year initiative designed to facilitate digital media platform observability at scale. The AIO integrates access to computational resources, data governance frameworks, and diverse data sourcing methods, ranging from platform-centric API-harvesting to user-focused data donations. The AIO addresses methodological challenges, ethics, the technical skills gap, and the complexities of regulatory compliance. We argue that shared infrastructures function as vital institutional interventions to ensure independent, reproducible, and ethical digital media research in a continuously evolving media environment.

Keywords

Australian Internet Observatory; data donation; digital media; observability; research infrastructure

1. Introduction

Digital platforms have become central infrastructures of communication, entertainment, education, employment, and political participation, shaping how information is produced, circulated, and encountered.

© 2026 by the author(s), licensed under a Creative Commons Attribution 4.0 International License (CC BY).

For communication and media researchers, this has generated unprecedented empirical opportunities and profound methodological challenges. While early research benefited from relatively open access to platform data and manageable data volumes, contemporary platform ecosystems are characterised by scale, heterogeneity, ephemerality, and opacity. Collecting and analysing such data at scale now requires advanced technical expertise covering data sourcing and processing, alongside resources to ensure rigorous governance. However, platforms' increasing power and the ways they mediate user-content interactions create significant access challenges: Relying solely on platforms for data access or to capture digital traces of user behaviour is no longer sufficient and often impossible. As Rieder and Hofmann (2020) argue, the ephemeral and powerful nature of algorithmic systems demands a new toolkit for comprehension and political oversight.

As a result, the capacity to systematically observe, analyse, and critique platform-mediated communication increasingly depends on the development of shared, open research infrastructures rather than isolated, project-level solutions. Driven by platformisation, datafication, and algorithmic governance, data access is now contingent on complex technical architectures with evolving access and use constraints. The methodological complexity is compounded by legal constraints; scholars must navigate uneven, context-dependent regulatory frameworks, varying censorship practices, and terms of service (ToS) that prioritise commercial goals and disregard academic inquiry. Declining application programming interfaces (APIs) accessibility and anti-scraping technologies present limitations that require additional expertise and infrastructure to address.

Such barriers threaten research sustainability and reproducibility while reinforcing epistemic inequalities, effectively marginalising scholars without specialised computational support. By pooling resources, shared infrastructures provide greater technical stability and governance support without overburdening individual researchers. This collective capacity enables rigorous platform observability grounded in open science principles: transparency, reproducibility, and accessibility (Dienlin et al., 2021).

In this article, we argue that a central challenge facing contemporary platform research is not simply the declining access to data, but the broader problem of observability: the capacity to systematically see, interpret, and evaluate how platform systems shape communication, visibility, and participation. Rather than treating data access as a technical bottleneck, we position observability as a methodological and institutional concern that determines what forms of platform power can be studied. We also argue for the development of inter-institutional and national-scale research infrastructures capable of supporting sustained, accountable platform observability.

We then present the Australian Internet Observatory (AIO; https://doi.org/10.25956/twvn-ca19) as a case study of a national research infrastructure (NRI) initiative across Australia. AIO aims to provide innovative new tools, governance frameworks, and methodological workflows that enable digital platform research at scale for humanities and social science researchers. AIO was established in 2024 with funding provided under the Australian Government's National Collaborative Research Infrastructure Strategy through investment by the Australian Research Data Commons and six university partner institutions: RMIT University, Queensland University of Technology, University of Queensland, University of Melbourne, Swinburne University, and Deakin University. AIO is funded from 2024 to 2028, in line with the Australian government's national research infrastructure roadmap, which is updated every five years (Department of

Education of the Australian Government, n.d.). This combination of government funding, institutional co-investment, and distributed governance ensures the structure is adaptable to fluctuations in funding and technological challenges.

The AIO case study provides findings from a national survey of researchers on the priority needs for digital platform research across the field. We outline the data collection methods and analytical pipelines under development, demonstrating how research infrastructure can function not merely as technical support but as a critical intervention.

2. General Challenges of Collecting Digital Platform Data

Communication and media research is increasingly shaped by a set of interconnected systemic forces that structure how online communication is observed and studied, including platformisation, ToS, technical barriers, and legal and governance arrangements.

2.1. Platformisation and the Changing Conditions of Media Research

Public communication has changed as the open web has become dominated by an increasingly limited number of digital platforms. They drive platformisation through the systematic transformation of communicative acts into quantified data (datafication) and algorithmic governance, turning public communication into a commodity contingent upon the platform and its technical architecture (Poell et al., 2019). Platformisation alters not only communicative practices but also the conditions under which they are studied. Platforms function as complex socio-technical systems that algorithmically organise visibility, relevance, and participation (Bruns, 2023). Communication is no longer concentrated in mass media but distributed across interconnected spaces, blurring the distinction between public and private interaction (boyd, 2010; Papacharissi, 2023). This fragmentation challenges conventional research, demanding resource-intensive methods to trace patterns across networks, modalities, and temporal scales.

Simultaneously, platforms shape communicative environments through content moderation, recommendations, and evolving functionality (Gongane et al., 2022). While automated and AI-driven technologies maximise engagement (Teepapal, 2025) and moderate harmful content (Gorwa et al., 2020), they also dictate what is amplified, suppressed, or transformed (Bandy, 2021; Sandvig et al., 2014). For researchers, this implies that platforms function not merely as venues but as algorithmic actors, increasing the complexity of empirical observation.

2.2. ToS, Commercial Considerations, and Asymmetric Power

ToS are a key governance mechanism that controls platform access, protects legal rights, and establishes binding guidelines for users (Brown et al., 2025). Beyond prohibiting automated data collection and restricting data reuse, ToS are platform-specific, ambiguous, and driven by commercial imperatives rather than academic needs (Fiesler et al., 2020; Hunter et al., 2018). Platforms have progressively restricted APIs, introduced paywalls, and deprecated access pathways, creating uncertainty regarding data sustainability (de Vreese & Tromble, 2023; Perriam et al., 2020).

This dynamic creates structural asymmetries between researchers and platforms. Scholars are expected to produce independent, critical knowledge about opaque systems operated by entities that are often unwilling to collaborate or hostile to researchers (Mir & Doctorow, 2021). Dependence on access mechanisms subject to unilateral withdrawal due to corporate or legislative shifts undermines reproducibility and long-term research data collection. Moreover, it favours well-resourced institutions able to absorb changes through technical workarounds or paid access, exacerbating existing inequalities within the field (Leeftink & Angus, 2023).

2.3. Technological Barriers to Data Collection and Analysis

APIs have historically been the primary tools for systematic, unobtrusive access to publicly visible platform data. However, the type, volume, and timespan of data are determined by each platform, often with varying access levels (Ohme et al., 2023). While major services, such as Twitter/X, Instagram, and Facebook, have restricted or withdrawn access to APIs in recent years, platforms such as Reddit, Mastodon, and Bluesky still facilitate data collection, enabling academic discovery.

As mainstream platform APIs declined, researchers have increasingly turned to ad hoc web scraping (Breuer et al., 2023; Freelon, 2018; Mancosu & Vegetti, 2020). However, newer dynamic web interfaces present technological challenges, preventing complete data capture by loading content only upon direct interaction (Luscombe et al., 2022; Turk et al., 2020). Tools that distinguish humans from machines, alongside personalised content delivery and rate limiting, make scraping fragile and resource-intensive. Maintaining functional data collection pipelines requires ongoing code engineering and technical expertise.

Simultaneously, platform communication is increasingly multimodal, combining text, images, and video in ways that resist single-method data analysis. While machine learning expands analytic possibilities, it also introduces dependencies on computational infrastructure (Grimmer & Stewart, 2013; van Atteveldt et al., 2021). Without shared tools and scalable environments, many researchers are effectively excluded from these methods.

2.4. Legal Regulation, Platform Governance, and Censorship

Platformisation evolves together with governance mechanisms, including regulatory frameworks and censorship (Halavais, 2019; Poell et al., 2019). While data protection laws address legitimate concerns regarding privacy and harm (Napoli, 2019), they also restrict data availability. Researchers face evolving compliance requirements affecting data collection, storage, and sharing, often with limited institutional support. Scholars must account for regulatory biases (Allen et al., 2021) and critically evaluate the suitability and implications of access methods (Brown et al., 2025). Embedding legal and ethical expertise within shared research infrastructure can mitigate these challenges by standardising governance and supporting researchers in navigating regulatory complexity (Glazunova et al., 2023).

On the other hand, EU legislation has also created new access opportunities. The General Data Protection Regulation mandates data ownership and portability, enabling users to export their digital traces as data download packages (DDPs). While offering rich data, DDPs are complex and often include multimodal files that require considerable resources for parsing and analysis. Using DDPs and deriving insights from data at scale requires coordinating across typically fragmented user bases and monitoring platform changes. Similarly, the EU Digital Services Act aims to increase transparency and accountability for very large online

platforms and search engines through researcher-focused obligations, though the practical implications remain unclear (Selling et al., 2025).

3. Methodological Challenges of Platform Research

Given these complexities, providing targeted research infrastructure requires multiple methods, including APIs, data scraping, and data donation, as well as tools for data cleaning, contextualisation, validation, and interpretation. To make valid empirical claims, researchers must understand data sources and the biases they introduce (boyd & Crawford, 2012; Ruths & Pfeffer, 2014). In some research contexts, achieving validity can be problematic, as platform data is not a direct reflection of reality (Ruths & Pfeffer, 2014) but rather comprises partial traces influenced by user dynamics, commercial incentives, and evolving algorithmic architectures (Bucher, 2018; Lazer et al., 2014).

3.1. Data Access and Bias

Platform-driven changes to APIs, moderation, and ranking introduce algorithmic confounding (Salganik, 2019). Given the opaque nature of major platforms, researchers risk drawing conclusions based on governance artefacts rather than underlying social processes. This confounding reinforces the need for shared infrastructure to stabilise data collection, track platform evolution, support cumulative knowledge, and foster critical debate on best methodological practices (Tromble, 2021).

Platform data interpretation is further complicated by overlapping biases (Tufekci, 2014): algorithmic bias shapes content visibility; reporting bias privileges specific engagement metrics (Rieder & Hofmann, 2020); and selection bias frequently skews samples toward highly active accounts (González-Bailón et al., 2014). These biases are compounded by cohort effects and uneven data availability, as shifting demographics, policies, or interfaces make cross-time and cross-platform comparisons methodologically fragile (Rogers, 2017). Unchecked, analyses risk reifying platform logic as social facts. Addressing this demands methodological reflexivity and infrastructural support to document sampling decisions, audit algorithmic conditions, and enable comparative analysis across datasets (van Atteveldt et al., 2021).

3.2. Beyond the "Notebook Paradigm"

Data access also raises reproducibility concerns. Computational communication researchers often advocate for a "Jupyter notebook" model that integrates data collection and analysis into shareable notebooks (van Atteveldt et al., 2022). This approach aligns with open science ideals by supporting transparency and explicit, replicable workflows (Dienlin et al., 2021). However, notebook-based workflows demand programming expertise, implicitly equating computational literacy with research proficiency. This creates barriers for non-coders, limiting broader inquiry (Schoch et al., 2024), particularly for qualitative or interpretive researchers, as notebook-centric approaches tend to favour quantitative analysis. While qualitative work is possible, notebook affordances prioritise tabular data and statistics over richer, contextual analyses (Kery et al., 2018). Thus, interactive visual environments enabling dynamic exploration remain historically under-supported within conventional workflows.

Recent developments in large language models (LLMs) complicate this landscape. By enabling researchers to generate and adapt code via natural language, LLMs lower entry barriers and decouple programming expertise from analytical capacity. However, LLM-generated outputs introduce new challenges: they can obscure analytical decisions, create subtle errors, and produce workflows that are difficult to document or reproduce, particularly when probabilistic.

These critiques do not reject notebooks and LLMs as valuable tools, but rather emphasise the risks of treating any single format as a universal solution. Research infrastructures must accommodate multiple modes of engagement such as codeless interfaces, visual exploration, and hybrid LLM-assisted workflows.

3.3. Technical Support and Skills Development

These limitations point to a broader challenge beyond the tools themselves: the skills and institutional capacities required for large-scale digital platform research. Researchers must increasingly engage with data engineering, system maintenance, documentation, version control, storage architectures, and long-term data stewardship. Though rarely visible in publications, this work is foundational to observability, reproducibility, and sustainable research programs. When informally absorbed by individuals or short-term projects, these responsibilities produce fragile infrastructures that are difficult to maintain or transfer.

LLM-based tools may ease this technical burden by supporting code generation, documentation, and exploratory analysis. They can also act as informal training supports for navigating unfamiliar technical workflows. However, these developments do not remove the need for underlying infrastructure; instead, they shift where expertise is required. Researchers must now address model reliability, prompt design, output validation, and the integration of LLM systems into established workflows.

At the same time, access to computational resources and technical expertise remains unevenly distributed across institutions and regions. Training in data engineering, software development, and advanced analytics is inconsistent across communication and media programs, reinforcing structural inequalities in participation in data-intensive research (D'Ignazio & Klein, 2023). As Schoch et al. (2024) argue, the increasing centrality of computational methods risks narrowing who can fully engage in data-driven research, especially when technical proficiency becomes an implicit prerequisite for scholarly legitimacy. These disparities are exacerbated by unequal access to high-performance computing, cloud infrastructure, and professional research software engineering support (Dutta et al., 2021).

Together, these challenges shape the need for training models that move beyond narrow computational upskilling toward infrastructural literacy. This involves understanding how platform data is generated, captured, stored, transformed, governed, and rendered observable across its lifecycle. It also encompasses data governance, ethics, documentation standards, collaborative workflows, and the institutional arrangements necessary to sustain long-term data collection amid platform volatility. Developing these capacities cannot be left to individual researchers or ad hoc teams. Instead, shared research infrastructures play a vital role by centralising technical expertise, providing supported environments, and embedding training within stable systems (Humphreys et al., 2021).

3.4. Research Infrastructure for Digital Platform Observability

Given the scale, multimodality, and material barriers of digital platforms, researchers require an infrastructure of commensurate scope. Moving beyond institution-level projects, we must invest in national-scale facilities to achieve observability and accountability reflecting the platforms themselves (Rieder & Hofmann, 2020). Such infrastructure must operate as a comprehensive system integrating computational resources, analytical tools, ethical governance, and support for diverse epistemic traditions.

Current platform research generally divides into platform-centric and user-centric approaches, each involving trade-offs regarding privacy, granularity, and researcher autonomy. Platform-centric methods use APIs or web scraping to facilitate large-scale retrospective analysis of public content, but leave scholars dependent on third-party pipelines. User-centric approaches rely on citizen science, i.e., collecting data via data donation using DDPs, mobile applications, or browser extensions (Ohme et al., 2023, p. 3), offering greater transparency and capturing fine-grained behaviours. However, they pose great privacy risks and face challenges in recruiting participants at scale.

Broadly speaking, method selection depends on the specific research problem, yet researchers universally require access to a diverse suite of tools. Given the complexity and resources involved, individual-level access is insufficient; successful data collection increasingly depends on infrastructure-level capacities. To provide an example of what such an infrastructure might look like, we examine the AIO as a case study. We explore the requirements, planning, and practical implementation of large-scale research infrastructure for media and communication, and the wider humanities, arts, and social sciences (HASS) fields.

4. Case Study: AIO

The need for AIO grew out of work undertaken at the Australian Research Council Centre of Excellence for Automated Decision-Making + Society (ADM+S; see https://www.admscentre.org.au), where researchers used browser extensions to collect data donations for research on advertising and search (see Angus, Obeid, et al., 2024; Bruns, 2022). AIO builds on this work with the aim of providing national research infrastructure available to Australian researchers across academia, government, and civil organisations, beyond the bespoke tools developed at the project level.

As shown in Figure 1, AIO comprises four pillars: the AIO platform, data collection tools, data governance and management, and outreach and training. The platform provides cloud-based workflows, data visualisation, and analysis tools. Data sourcing includes multimodal content via data donation, APIs, and LLM-generated synthetic data. The governance and management pillar establishes secure architectures and ethical frameworks adhering to FAIR (findable, accessible, interoperable, reusable) and CARE (collective benefit, authority to control, responsibility, ethics) principles, assisting researchers with institutional compliance. The outreach and training stream builds capability through workshops, materials, and research engagement.

Access to the platform is managed via secure authentication (e.g., CILogon), and prior Human Research Ethics Committee approval is required for researchers to conduct data collection. Research groups retain ownership of collected data and determine third-party access levels. A self-service interface launches in August 2026 for

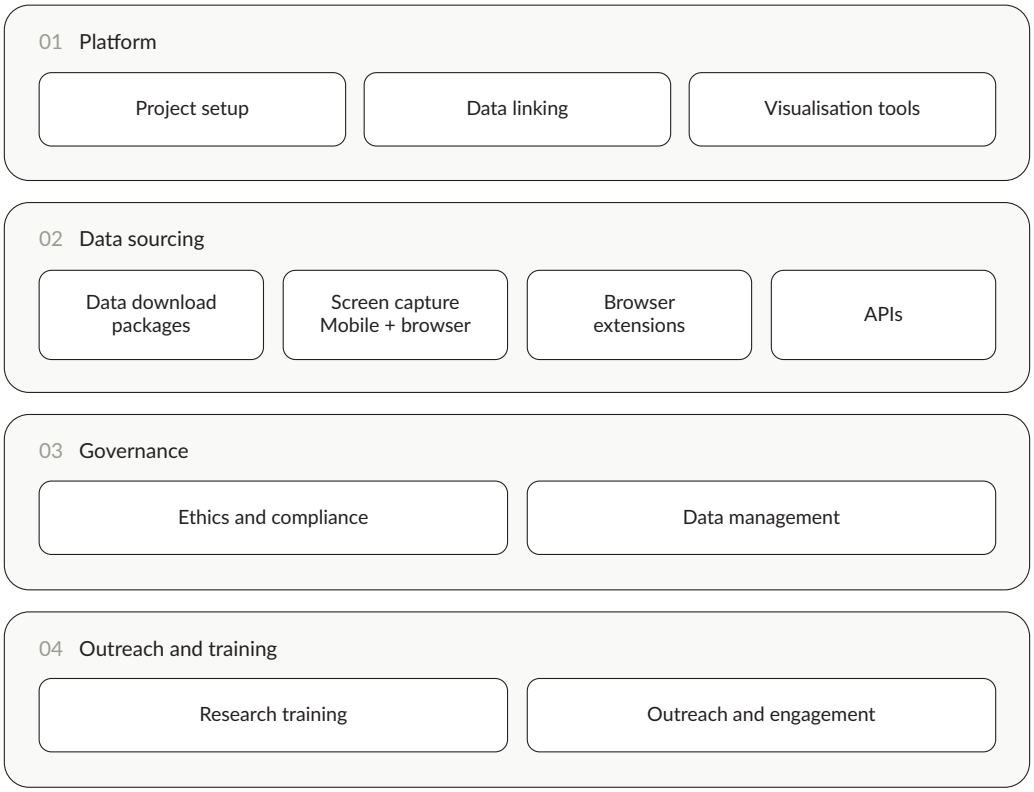

Figure 1. Infrastructure-level components of the AIO.

smaller projects requiring no additional configuration and minimal support. At the same time, the team also collaborates with select large projects to co-design functionality based on user surveys and the project plan, subsequently making these features available to all researchers.

As a federally funded national research infrastructure, AIO aims to support as many projects as possible within team capacity. Pending second-phase funding in 2028, a sustainability model with minimal cost recovery will be considered. While AIO's funding is Australia-specific, its open-source code, tools, and methods are globally adaptable; the team also collaborates with observatories across Europe and the US via international conferences and the Data Donation Institutional Network (see https://datadonation.eu/community) to share expertise in developing research facilities (see also https://internetobservatory.org.au/networks).

4.1. Research Needs Survey

To ground the development of the AIO in the needs of the research community, a national survey was conducted from July to December 2025 to identify researchers' practices, challenges, and preferences. Findings validated design assumptions and development priorities while addressing gaps in planned capabilities. For example, a strong interest in data donation tools, visual interfaces, and ethical governance directly shaped the project's strategy.

The survey consisted of 24 questions across four sections, using both multiple-choice and open-ended items. These sections covered demographic and professional information, research interests and data needs

(including platforms and collection methods), ethical concerns and analytical requirements (including generative AI), and training preferences.

The survey was distributed via the AIO website, partner organisations, professional networks, and conferences. Using snowball sampling, participants were also encouraged to share the link within their networks. Of the 105 responses received, the results are interpreted as indicative of the priorities facing scholars reached by the AIO, rather than representative of the entire social science field. However, the diversity of respondents across Australian sectors, disciplines, and career stages provides insight into the overall landscape of research needs and interests.

Primarily from the Australian higher education sector (81%), the sample also included not-for-profits (7%), government, and business (11%). Respondents were largely academics (62%), followed by students (15%) and academic support professionals (12%). Disciplinary representation was heavily weighted toward HASS, with the largest cohorts in language, communication, and culture (48%) and human society (42%), as shown in Figure 2. Category definitions followed the Australian and New Zealand Standard Research Classification; therefore, the human society category included fields such as political science, sociology, criminology, social work, and policy studies.

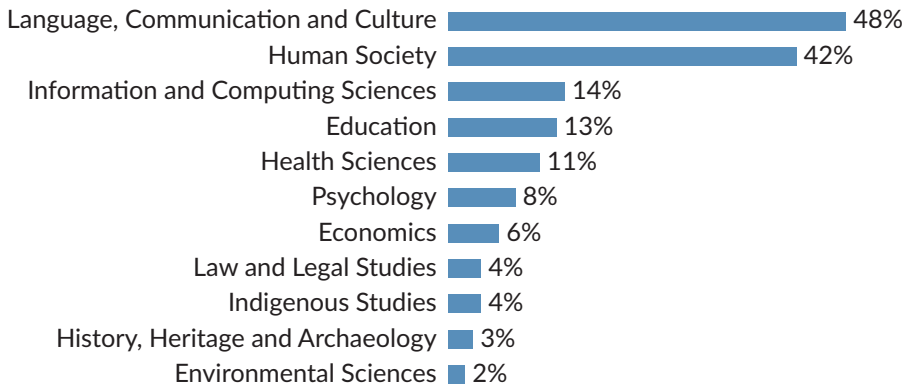

Figure 2. Survey participants' reported areas of research.

Respondents prioritised complex, algorithmically driven platforms, with Instagram attracting the most interest (58%), followed by TikTok (52%), YouTube (51%), and Facebook/Meta (49%), as shown in Figure 3. Generative AI was also a major interest for 44% of participants, while around a third targeted news sites, messaging apps, and search engines. These results reflect the importance of visual and short-form video, reinforcing the need for multimodal content. Even though text remained the preferred data format (73%), significant interest in images (61%), metadata (54%), and video (50%) was reported. The high metadata demand suggests a focus on platform architectures and algorithmic processes alongside traditional content analysis. Platform interest also varied by discipline, with generative AI attracting higher interest among education (62%) and computing sciences (54%) scholars than among language, communication, and culture (33%) and human society (37%) researchers.

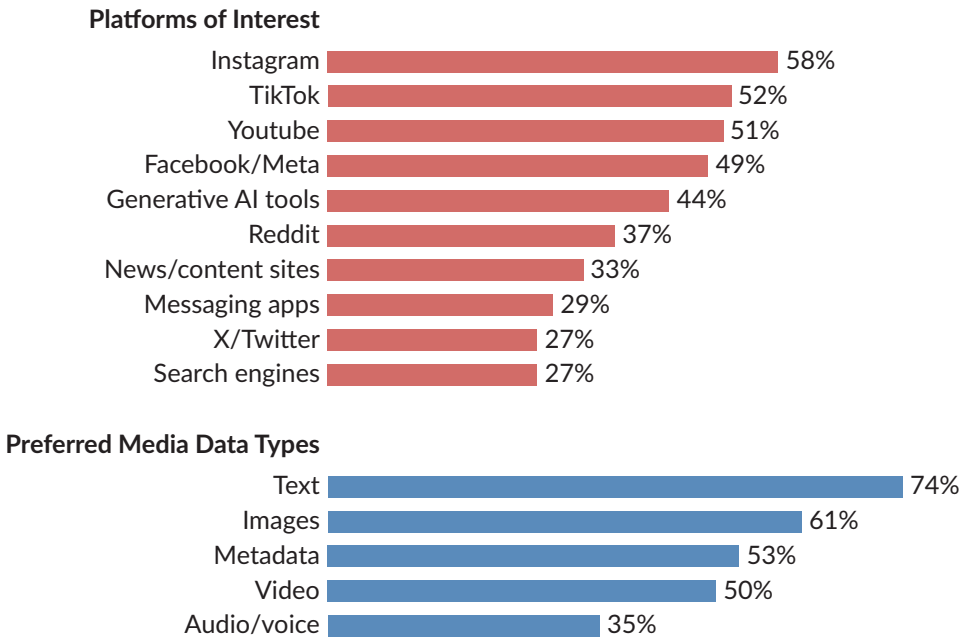

Figure 3. Reported platform and methods of interest among survey participants.

Figure 4 shows interest in data collection methods by self-reported experience level. Respondents selected multiple options, reflecting their research interest. Web scraping (55%), data donations (51%), and APIs (47%) were the most frequently selected methods. Web scraping interest remained consistent across groups. However, only 36% of beginners selected APIs, which was significantly lower than experienced (55%) or expert (53%) respondents; it likely reflects technical barriers or declining API availability. At the same time, data donation appealed to beginners (50%) and those with experience (55%), but less so to experts (40%), suggesting that it holds particular appeal for researchers seeking lower-barrier routes to participant data. Although APIs have been among the preferred methods of data access, their positioning below web scraping and data donation may reflect that API-based methods alone are insufficient or that the main platforms of interest no longer offer APIs for data access.

Data access is an analytical as well as technical challenge; many researchers prioritise tools for data interpretation over raw computational power. Content analysis (68%) and annotation (58%) tools were most prioritised, while roughly a third preferred network analysis (35%) and machine vision (38%). These responses articulate a demand for interactive environments that would allow researchers to classify, annotate, and make sense out of unstructured datasets.

Ethics and governance concerns raised by the respondents also align with the complex regulatory landscape of platform research. Participant privacy was the most cited issue (71%), followed by data security and management (59%). One-third of researchers explicitly stated "legal challenges from digital platforms" as a barrier, confirming that restrictive ToS influence research design. These responses indicate that normative compliance is a significant burden for individual researchers, potentially discouraging engagement with certain platforms. Simultaneously, a segment of researchers (11%) is concerned with Indigenous data governance, which also points to the importance of the FAIR and CARE data principles. Such frameworks require institutional intervention to ensure that data involving Indigenous communities is managed with appropriate cultural protocols.

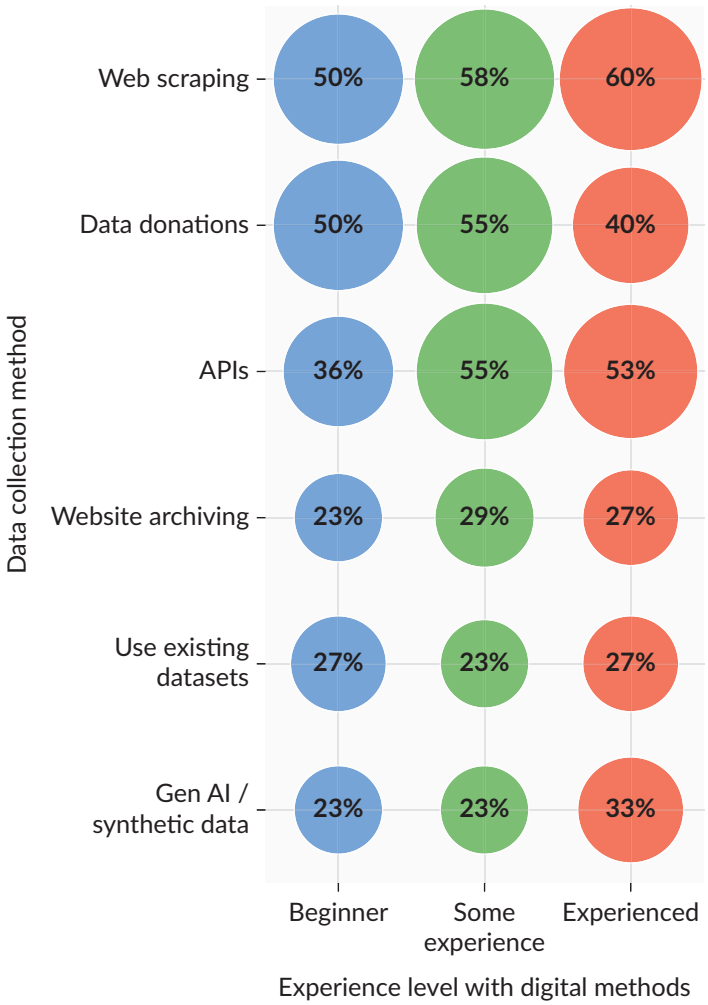

Figure 4. Data collection method interest by experience level.

Open-text responses showed diverse concerns across communication and social research. The overarching theme was the opacity of platform governance (the algorithmic "black box") and its broader societal impacts. Specific research topics included targeted advertising and racial discrimination, online harms, medical and financial scam ads, pro-anorexia content, tech-facilitated gender-based violence, and state-backed information operations. Respondents also prioritised personalised delivery and platform processes that account for specific demographics, such as teenagers and the elderly, who see and interact with information differently based on their identities and behavioural patterns.

When asked about training needs, 88% of researchers identified "data collection tools and methods" as their primary area of interest, outpacing data analysis (64%) and visualisation (54%), as shown in Figure 5. There remains a perception that digital research is largely driven by data access, and, simultaneously, it remains the largest bottleneck for many scholars. Online workshops (54%) were the preferred format for training, with online courses or videos being the preference for 43%, although 57% pointed out that the mode of delivery largely depended on the subject.

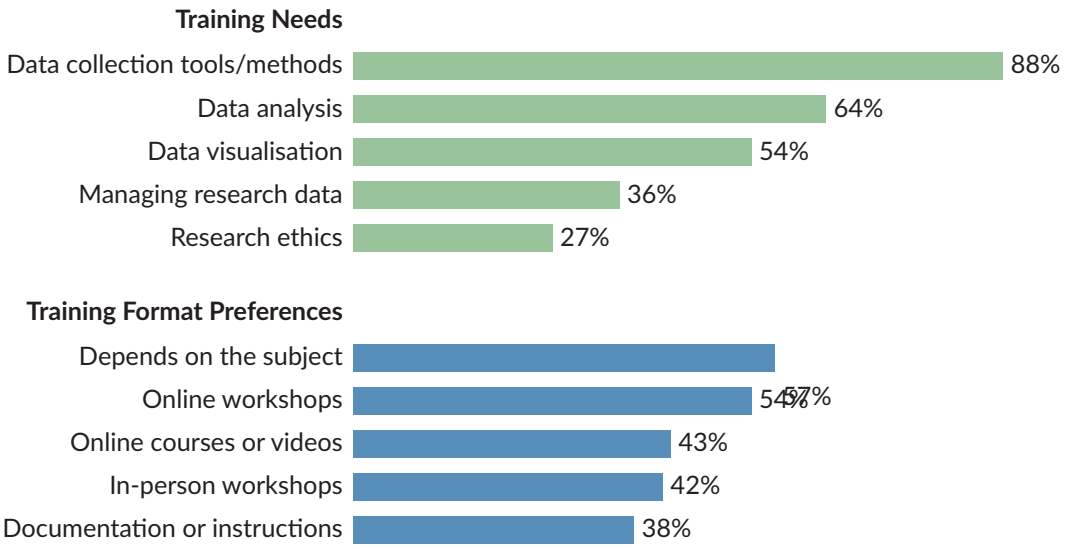

Figure 5. Reported research needs, training preferences, and uses of AI in research.

The survey data support the idea that different communities of researchers within HASS disciplines are eager to engage in research on digital media but face a variety of hurdles that prevent them from starting new projects. Limited data access, the lack of tools suitable for scholars without technical expertise, and limitations imposed by platforms and legislation create barriers that demonstrate the need for better ecosystems that integrate user-centric research, multimodal analysis, and rigorous ethical governance. The following sections describe how the AIO addresses these challenges through its components as an infrastructure: data donation tools, which responds directly to the respondents seeking user-centric platform data analysis; API-based data collection, addressing the need for platform-centric analysis given the interest in APIs; ethics and data governance, addressing the privacy concerns and legal challenges; and the research training program, designed to address the skills gap.

4.2. AIO Data Donation Tools and Methods

A key AIO data collection approach involves supporting data donation methods, including DDPs and screen capture on mobile or in a browser. Each workflow targets specific types of user-centred data collection and enables different modes of observation.

DDPs are a standardised format for collecting files that users can export from platforms under data portability and privacy rights, such as the General Data Protection Regulation. Building on existing European frameworks such as port (Boeschoten et al., 2023) and data donation module (Pfiffner et al., 2024), the AIO workflow prioritises data donors' privacy. As presented in Figure 6, it processes data locally on participants' devices using browser-side scripts, filtering out sensitive information before uploading the dataset to a secure environment where the researcher can access it. Designed in this way, the DDP system provides a secure pipeline and removes the technical burden of data donations on both participant and researcher sides.

Screen capture is another data donation method that takes a different approach, using mobile apps and browser extensions to capture data from participants while they are online. Mobile interaction data has long been one of the most elusive forms of digital communication data. Unlike web-based content, mobile media

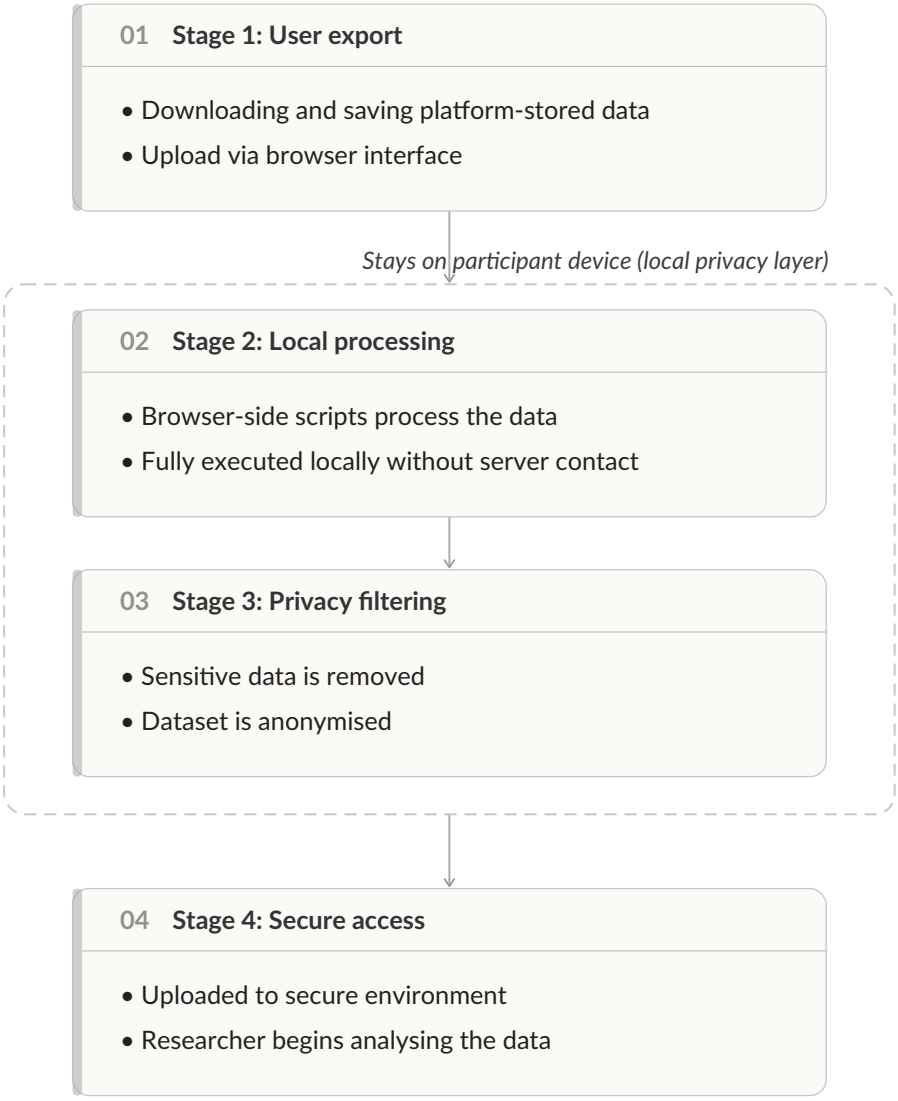

Figure 6. AIO DDP data donation process. Note: The process is step-by-step and begins with the respondent downloading their data from a social media platform, accessing the data donation system, and selecting their files.

is largely inaccessible through public interfaces, APIs, or archival tools, despite mobile devices being the primary site for everyday platform use. As a result, critical dimensions of contemporary communication, including personalised feeds, ephemeral content, in-app advertising, and algorithmically sequenced recommendations, have remained largely unobservable with conventional data-collection approaches.

To address this, the AIO has developed a general-purpose mobile screen capture framework: a purpose-built mobile app that runs passively in the background of participants' phones and records images, text, and video as they appear on participants' phone screens across major platforms including Facebook, TikTok, and Instagram. By operating at the level of the delivered interface rather than the platform-provided endpoints, the framework enables the capture of real-time interaction data as users experience it. Importantly, the framework is modular and can be configured for different research purposes, enabling targeted capture of specific content types.

One such implementation is the AIO Mobile Observation Ad Toolkit (MOAT), which operationalises the framework for studying mobile advertising (Angus, Hayden, et al., 2024; Obeid et al., 2026). In contrast to general capture approaches that collect broad streams of user activity for later filtering, MOAT is purposefully tuned to detect and extract advertising content at the point of capture. This design ensures that only data directly relevant to the research question, advertisements themselves, is retained and transmitted from the participant's device, substantially reducing the collection of extraneous or sensitive material. False positives are rare (approximately one in every 10,000 ads donated) and are subsequently removed through a secure, researcher-mediated validation process.

Using on-device screen capture, AIO MOAT allows researchers to study platform-based advertising as it is rendered on participants' screens, rather than as it is selectively made available through platform-controlled access mechanisms. This provides visibility into features often absent from other forms of platform data, including the sequencing and repetition of content delivery, the timing of exposure, and "bursty" patterns of advertising distribution that are difficult to detect through aggregate metrics alone.

The workflow, depicted in Figure 7, supports comparative analysis of how content is targeted and delivered across demographic groups, user clusters, and browsing conditions. By capturing mobile interaction data in situ, the framework responds directly to the methodological challenges outlined earlier, demonstrating how research infrastructure can extend observability into domains where data access is otherwise technically restricted, commercially protected, or entirely unavailable.

Similar to the mobile app, browser-based screen capture enables participant-centred data donation during everyday online activity. Extending open-source approaches developed by AlgorithmWatch and the ADM+S Australian Search Experience (Bruns, 2022), the AIO browser extension operates as a customised plug-in architecture that runs background processes during normal browsing to capture web page data, which is then submitted to a secure, cloud-hosted database. The system is designed as a generalised framework that incorporates a streamlined cloud pipeline and a researcher-facing interface for data management. This allows researchers to configure and deploy the extension within their own studies, collecting web and platform data alongside voluntarily supplied demographic information, without the need to engineer bespoke backend infrastructure from scratch.

To extend their analytical value, screen capture infrastructures are typically deployed alongside complementary methods, most commonly surveys and questionnaires. While capture tools provide high-resolution data on the structure and sequencing of content delivered to users, surveys gather unobservable contextual information such as demographics, attitudes, and self-reported behaviours. Linking observed exposure with self-reported perceptions supports richer analysis. These tools also feed into secure, researcher-managed dashboards that, in some study designs, allow participants to review and reflect on their donated data, facilitating participant validation and engagement.

This combined approach allows researchers to examine how content is experienced across user groups, how it aligns with participants' own accounts of their media use, and how exposure patterns relate to broader attitudinal or behavioural outcomes. Participant-facing dashboards further support collaborative interpretation, enabling users to contextualise and interrogate their data alongside researchers. This adds a layer of validation and reflexivity, situating observed platform outputs within participants' lived contexts.

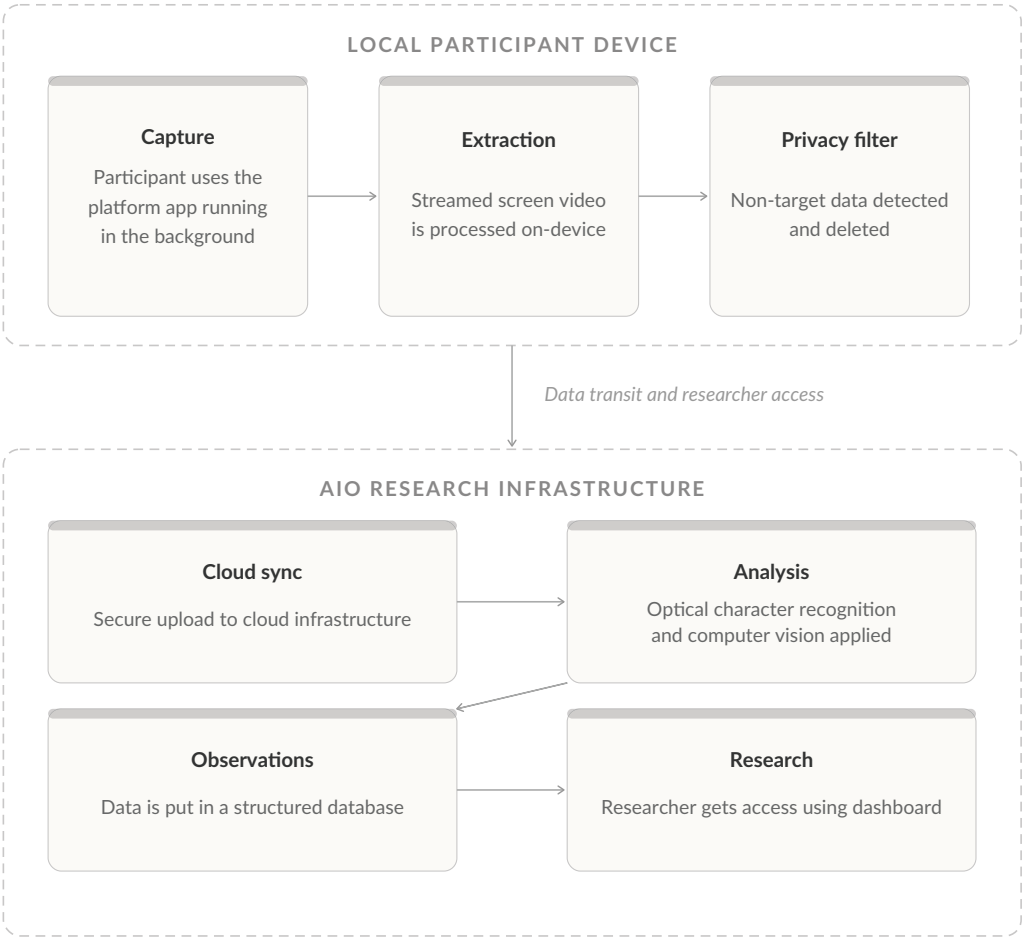

Figure 7. AIO mobile screen capture tool for data collection and processing workflow. Notes: The data is stored on the participant's device, where it is processed and filtered; then, upon the participant's examination and confirmation, it is uploaded to the secure infrastructure.

4.3. API Data Collection Tools

As the survey results show, researchers are interested in both platform-centric and user-centric approaches, including the use of APIs where possible. To address the challenges of large-scale multimodal data collection, the AIO is developing API-based data capture and visualisation tools through the AIReD (Australian Internet Observatory Research Dashboard) service, which harvests data from BlueSky, Flickr, Mastodon, Reddit, and YouTube. To incorporate news media, AIReD connects to GDELT, a database of electronic and print news reports, and maintains archived Twitter (X) data from before 2023 restrictions. Together, these resources facilitate longitudinal, comparative, and historical research.

AIReD prioritises Australia-related content by targeting local Reddit groups, the Australian Mastodon server (see https://mastodon.au), and regional harvesters. Only public posts are collected from the Mastodon.au node, consistent with the instance's own privacy policy, which acknowledges public content as publicly accessible, consistent with emerging norms regarding Mastodon data for research (Wähner et al., 2024). The repository currently holds over 484 million posts, each cleaned and processed through algorithms for topic modelling and sentiment analysis. Authentication via CILogon ensures equitable access for all researchers at Australian academic institutions without requiring additional registration.

The tool integrates adjustable graph, temporal, and statistical visualisations of aggregate data. It enables users to explore trending topics, search across datasets, and perform natural language queries using LLMs for instant analysis. To understand what is happening on a given day across the Australian social media landscape, topic modelling is used. The platform generates topic network graphs where nodes represent term clusters and edges indicate links between them. Users can adjust the relatedness threshold based on the number of shared terms required to form an edge; the maximum is 30, as higher values yield few connections. Figure 8 illustrates a BlueSky topic modelling graph from December 2025, identifying the war in Ukraine as a key topic. In these visualisations, node radii are proportional to the post volume within each cluster. Topic modelling provides an overview of the daily Australian social media landscape, identifying sets of terms that repeat as conversations evolve.

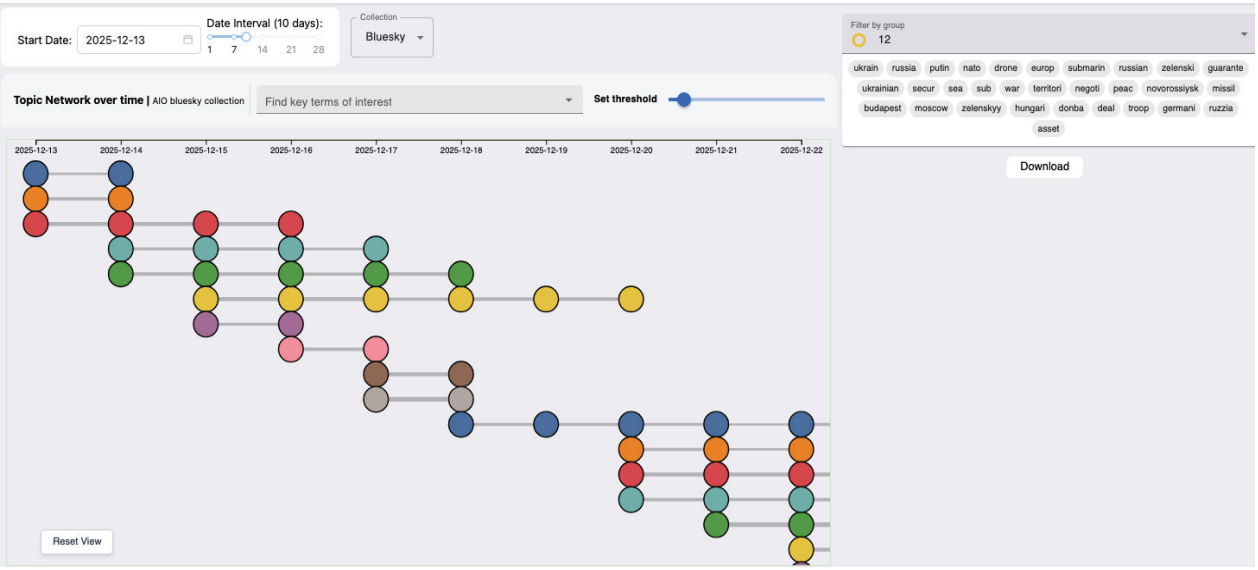

Figure 8. AIReD topic modelling tool. Note: The interface allows for time interval, word count threshold, and group selection to display corresponding posts.

The AIReD dashboard also supports natural language querying of the content. Users can then zoom in on data sets that may be of interest using typical LLM conversations within a given user (ChatGPT-like) session, e.g., how many posts have been made on the Russia–Ukraine war across different platforms and what their sentiment is. Such queries can also highlight differences among user communities and how they are handled by the different platforms.

An API is also offered to dynamically query and use the AIReD data collections. Example Jupyter notebooks using data from the AIReD API demonstrate how the data collections offer insights from data at scale. The overall data flow associated with the AIReD environment is shown in Figure 9.

It is important to note that users cannot directly download the raw harvested data from AIReD, as this would violate the ToS of the various platforms. Instead, users can download the identifiers/URLs for the data that is of interest. They can then directly acquire (rehydrate) the data using the post IDs (Sinnott et al., 2023). Examples of how to achieve this are available to users through Jupyter notebooks and training materials.

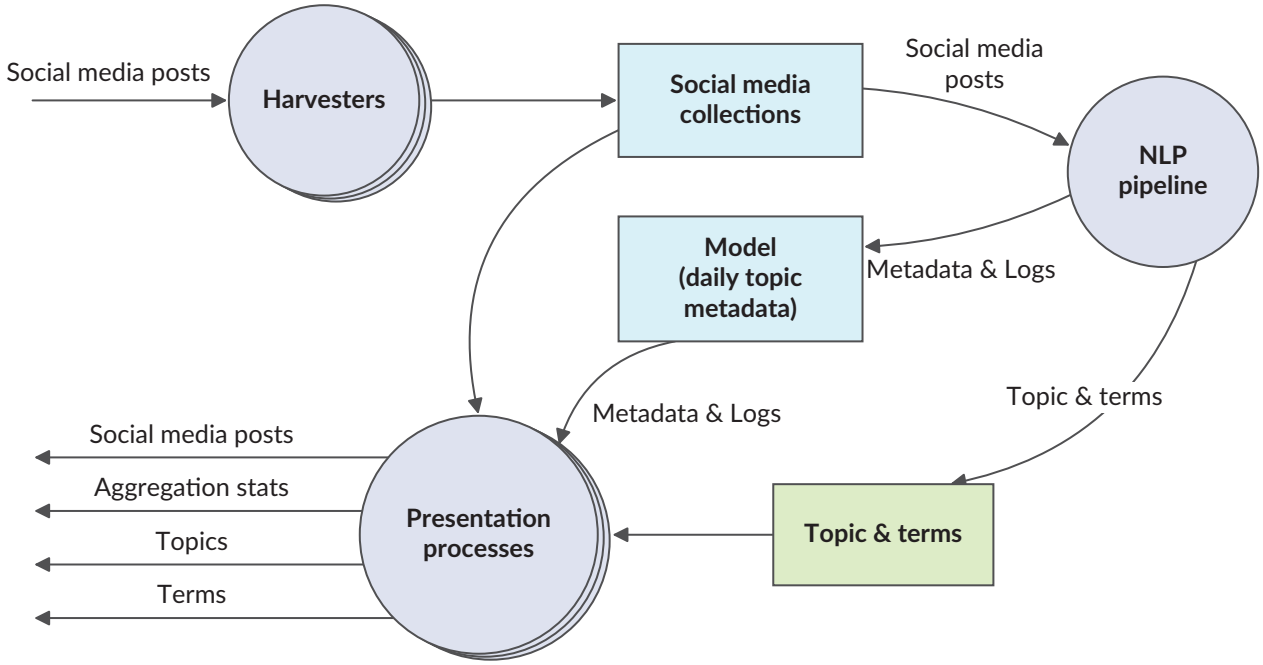

Figure 9. AIReD environment data flow. Notes: Social media posts are collected by platform-specific harvesters that then form social media collections, either presented directly or processed through the natural language processing (NLP) pipeline.

4.4. Ethics and Data Governance

AIO addresses uneven regulatory frameworks, blurred public-private distinctions, and epistemic inequalities through integrated ethical governance. Aligned with the National Statement on Ethical Conduct in Human Research 2025, the Australian Code for the Responsible Conduct of Research 2018, and the Privacy Act 1988, AIO serves as a centralised infrastructure. In this way, it shifts the burden of legal interpretation from individual scholars to a standardised baseline for compliance, helping researchers navigate inconsistent platform ToS.

The governance framework provides a clear pathway for data access that satisfies both national standards and legislative requirements. AIO employs a trusted research environment based on the "five safes" framework: safe projects, people, data, settings, and outputs. Supporting this, the infrastructure incorporates role-based access controls via federated identity management (CILogon), secure cloud workspaces, and metadata-driven audit trails. CILogon facilitates access across institutional affiliations, while storage is managed to rigorous standards, enabling sensitive data protection and addressing the complexities of maintaining multiple large-scale datasets.

For public and semi-public data collected via APIs, the AIO provides baseline criteria for determining "reasonable expectations" of privacy, helping researchers justify consent waivers where appropriate. For user-centric tools involving DDPs and screen capture, the AIO interface allows participants to review, pause, or withdraw their contributions at any time, thus respecting participant autonomy. Additionally, the governance framework in development is designed to align with the FAIR and CARE principles, ensuring that research involving Aboriginal and Torres Strait Islander peoples is guided by community authority and benefit-sharing. The framework also includes specific risk assessment protocols for research involving young

people and culturally diverse communities, aimed at supporting diverse epistemic traditions and preventing the amplification of algorithmic harms against marginalised groups.

Recognising that human research ethics committees often struggle to evaluate the novel risks associated with digital methods, the AIO is building a library of standardised resources. These include pro forma documents, risk assessment matrices, and technical descriptions of tools and methods. These resources translate complex technical workflows into the language of institutional governance, helping researchers navigate ethics approvals and ensuring consistency across universities.

4.5. AIO Research Training Program

The lack of technical knowledge and the risk of marginalising researchers who do not use specialised computational expertise must be addressed by providing training opportunities, assistance, and resources that are highly accessible, regardless of field. The AIO's approach in this regard is through the development of the research training program that serves as an instrument to bring computational and non-computational scholars together through interactive online resources, video and textual materials, workshops, summer schools, and collaboration/coordination with projects that are interested in using digital methods but require additional assistance in understanding and adapting them to their research design.

The AIO research training program bridges the skills gap through a hybrid model. It aims to combine a self-paced online learning system for foundational skills with in-person workshops for project-specific methodological support. This dual approach aims to upskill HASS researchers in both conceptual framing and technical execution. The program's design focuses on developing online materials as part of a permanent, self-paced online learning system that includes video modules and "how-to" guides. The modules serve as the foundation for in-person initiatives, including events developed in partnership with other research centres. It is designed so that interested researchers can acquire foundational skills and knowledge about the specific AIO tools at their own pace, while project-specific designs and methodological challenges are addressed with hands-on support as needed.

5. Conclusion

The contemporary platform ecosystem, characterised by scale, ephemerality, and opacity, demands a response that goes beyond isolated project-level solutions. Rather than treating data access as a mere technical bottleneck, we argue that it is a methodological and institutional challenge (Tromble, 2021). Continuous platformisation and rapidly changing media environments, complex legal issues, barriers to accessing computational resources and methods, specialised skill shortage, and the necessity of maintaining high ethical standards and research rigour cannot be addressed by individual researchers and teams. We argue that an infrastructure-level solution is necessary to provide sustained, accountable platform observability (Rieder & Hofmann, 2020).

The AIO as an infrastructure project offers an integrated approach to resolve these systemic challenges. The platformisation-driven decline of API access and the technical barriers are addressed by diverse, user- and platform-centric data collection strategies including API-based data harvesting, DDP workflows, and screen capture tools (Angus, Hayden, et al., 2024; Obeid et al., 2026). Ethical, data management, and

compliance limitations are mitigated through the standardised governance framework, centralised ethics compliance resources, and the five safes principle, as well as tool functionality, redistributing the interpretive burden away from individual scholars. The notebook paradigm and the skills gap are further addressed in the development of visual interfaces, accessible documentation, training, and other elements of the infrastructure program. Together, the AIO components represent national research infrastructure that enables observability and is aligned with the open science principles of transparency, reproducibility, and accessibility (Dienlin et al., 2021).

However, there are limitations to what the AIO can potentially resolve. The infrastructure is Australia-specific in its scope, funding, governance, and regulatory alignment. Replicating the AIO elsewhere would depend on institutional investment and sustained government commitment. At present and in the foreseeable future, platform coverage will remain limited; it remains challenging to access data from the platforms of greatest interest among surveyed researchers (e.g., Instagram, TikTok, and Facebook). User-centric methods of access are contingent on participant recruitment and remaining up-to-date with changes to platform data schemas. Even existing pipelines remain contingent on platform cooperation, making dashboards such as AIReD reactive to platform decisions. The sustainability model for such infrastructures is also under development and may include the implementation of cost-recovery mechanisms, which might then reintroduce the very access inequalities the AIO was established to address. While cost-recovery models may make projects more feasible in the long run, national infrastructure models need to be held accountable to the original principles they are built upon.

Internationally, AIO is not alone, and a growing number of infrastructures are working toward platform observability across different national contexts and disciplines, representing alternative forms of funding, structure, and focus. What distinguishes AIO is the combination of a nationally funded institutional status, a multi-method infrastructure integrating platform-centric and user-centric approaches, data governance guided by FAIR and CARE principles, and a training program designed to facilitate participation. AIO draws from and extends existing international standards, one key example being the port framework (Boeschoten et al., 2023), developed in the Netherlands and adopted across European research communities. Port provides an open-source pipeline for DDP collection, and AIO builds on this by integrating it within an approach that combines DDP, mobile screen capture, and browser extensions.

Developing infrastructures is only a single step toward independent and sustainable research. Since data access remains contingent on evolving regulatory frameworks and platform algorithms, the technical workarounds currently implemented are inherently reactive. Long-term successes require advocacy for structural and legal changes that facilitate platform transparency. Continued investment and institutional engagement are critical to maintaining these complex systems.

Because digital platforms operate globally, establishing global infrastructure through collaboration with emerging counterparts worldwide will create a network in which tools, governance protocols, and educational resources are shared toward common goals rather than duplicated. Involving more local institutions and the international community in the development of standardised approaches and practices, such as data donation and ethical frameworks, will be conducive to generating better knowledge about our societies and addressing ongoing problems of platform accountability around the world.

Acknowledgments

The Australian Internet Observatory (https://doi.org/10.25956/twvn-ca19) is a co-investment partnership with the Australian Research Data Commons (ARDC) through the HASS and Indigenous RDC (https://doi.org/10.3565/hjrp-b141). The ARDC is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS).

Funding

Publication of this article in open access was made possible through the institutional membership agreement between the University of Melbourne and Cogitatio Press.

Conflict of Interests

Bogdan Mamaev and Amanda Lawrence have been employed by the Australian Internet Observatory in research and infrastructure development roles relevant to the work described in this article. Daniel Angus and Richard Sinnott are chief investigators of the Australian Internet Observatory and have leadership responsibilities relating to its strategic direction, governance, and research program. No other conflicts of interest are declared.

LLMs Disclosure

Grammarly was used for spelling and grammar correction where appropriate. For word count reduction, Gemma 4 31B IT was used to identify sentence structures that could be simplified, and some of these suggestions were manually applied to longer paragraphs and sentences.

References

Allen, J., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Research note: Examining potential bias in large-scale censored data. *Harvard Kennedy School Misinformation Review*, *4*(2). https://doi.org/10.37016/mr-2020-74

Angus, D., Hayden, L., Obeid, A. K., Tan, X. Y., Carah, N., Burgess, J., Parker, C., Andrejevic, M., Fordyce, R., Cellard, L., & Bagnara, J. (2024). Computational methods for improving the observability of platform-based advertising. *Journal of Advertising*, *53*(5), 661–680. https://doi.org/10.1080/00913367.2024.2394156

Angus, D., Obeid, A., Burgess, J., Parker, C., Andrejevic, M., Bagnara, J., Carah, N., Fordyce, R., Hayden, L., Lewis, K., O'Neill, C., Albarrárran-Torres, C., & Cellard, L. (2024). *The Australian ad observatory: Technical and data report* (Working paper 009). ARC Centre of Excellence for Automated Decision-Making and Society. https://eprints.qut.edu.au/245657

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), Article 74. https://doi.org/10.1145/3449148

Boeschoten, L., de Schipper, N. C., Mendrik, A. M., van der Veen, E., Struminskaya, B., Janssen, H., & Araujo, T. (2023). Port: A software tool for digital data donation. *Journal of Open Source Software*, *8*(90), Article 5596. https://doi.org/10.21105/joss.05596

boyd, d. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), *A networked self: Identity, community, and culture on social network sites* (pp. 47–66). Routledge. https://doi.org/10.4324/9780203876527-8

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Breuer, J., Kmetty, Z., Haim, M., & Stier, S. (2023). User-centric approaches for collecting Facebook data in the 'post-API age': Experiences from two studies and recommendations for future research. *Information, Communication & Society*, *26*(14), 2649–2668. https://doi.org/10.1080/1369118X.2022.2097015

Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., & Zimmer, M. (2025). Web scraping for research: Legal, ethical, institutional, and scientific considerations. *Big Data & Society*, *12*(4). https://doi.org/10.1177/20539517251381686

Bruns, A. (2022). *Australian search experience project: Background paper* (Working Paper No. 1). ARC Centre of Excellence for Automated Decision-Making and Society. https://doi.org/10.25916/k7py-t320

Bruns, A. (2023). From "the" public sphere to a network of publics: Towards an empirically founded model of contemporary public communication spaces. *Communication Theory*, *33*(2), 70–81. https://doi.org/10.1093/ct/qtad007

Bucher, T. (2018). *If…then: Algorithmic power and politics*. Oxford University Press.

Department of Education of the Australian Government. (n.d.). *National research infrastructure*. https://www.education.gov.au/national-research-infrastructure

de Vreese, C., & Tromble, R. (2023). The data abyss: How lack of data access leaves research and society in the dark. *Political Communication*, *40*(3), 356–360. https://doi.org/10.1080/10584609.2023.2207488

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D. A., Vermeulen, I., Fisher, J. T., Banks, J., Weber, R., Ellis, D. A., . . . de Vreese, C. H. (2021). An agenda for open science in communication. *Journal of Communication*, *71*(1), 1–26. https://doi.org/10.1093/joc/jqz052

D'Ignazio, C., & Klein, L. F. (2023). *Data feminism*. MIT Press.

Dutta, M., Ramasubramanian, S., Barrett, M., Elers, C., Sarwatay, D., Raghunath, P., Kaur, S., Dutta, D., Jayan, P., Rahman, M., Tallam, E., Roy, S., Falnikar, A., Johnson, G. M., Mandal, I., Dutta, U., Basnyat, I., Soriano, C., Pavarala, V., . . . Zapata, D. (2021). Decolonizing open science: Southern interventions. *Journal of Communication*, *71*(5), 803–826. https://doi.org/10.1093/joc/jqab027

Fiesler, C., Beard, N., & Keegan, B. C. (2020). No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 187–196. https://doi.org/10.1609/icwsm.v14i1.7290

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Glazunova, S., Ryzhova, A., Bruns, A., Montaña-Niño, S. X., Beseler, A., & Dehghan, E. (2023). A platform policy implementation audit of actions against Russia's state-controlled media. *Internet Policy Review*, *12*(2). https://doi.org/10.14763/2023.2.1711

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: Current status and future directions. *Social Network Analysis and Mining*, *12*(1), Article 129. https://doi.org/10.1007/s13278-022-00951-3

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, *38*, 16–27. https://doi.org/10.1016/j.socnet.2014.01.004

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, *22*(11), 1567–1581. https://doi.org/10.1080/1369118X.2019.1627386

Humphreys, L., Lewis, N. A., Jr., Sender, K., & Won, A. S. (2021). Integrating qualitative methods and open science: Five principles for more trustworthy research. *Journal of Communication*, *71*(5), 855–874. https://doi.org/10.1093/joc/jqab026

Hunter, R. F., Gough, A., O'Kane, N., McKeown, G., Fitzpatrick, A., Walker, T., McKinley, M., Lee, M., & Kee, F. (2018). Ethical issues in social media research for public health. *American Journal of Public Health*, *108*(3), 343–348. https://doi.org/10.2105/AJPH.2017.304249

Kery, M. B., Radensky, M., Arya, M., John, B. E., & Myers, B. A. (2018). The story in the notebook: Exploratory data science using a literate programming tool. In R. Mandryk & M. Hancock (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). ACM. https://doi.org/10.1145/3173574.3173748

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Leeftink, D., & Angus, D. (2023). The research stack: A framework for data-driven humanities and social science. In W. Housley, A. Edwards, R. Beneito-Montagut, & R. Fitzgerald (Eds.), *The Sage handbook of digital society* (pp. 131–147). Sage. https://doi.org/10.4135/9781529783193.n9

Luscombe, A., Dick, K., & Walby, K. (2022). Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity*, *56*(3), 1023–1044. https://doi.org/10.1007/s11135-021-01164-0

Mancosu, M., & Vegetti, F. (2020). What you can scrape and what is right to scrape: A proposal for a tool to collect public Facebook data. *Social Media + Society*, *6*(3). https://doi.org/10.1177/2056305120940703

Mir, R., & Doctorow, C. (2021). *Facebook's attack on research is everyone's problem*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2021/08/facebooks-attack-research-everyones-problem

Napoli, P. M. (2019). *Social media and the public interest: Media regulation in the disinformation age*. Columbia University Press.

Obeid, A., Angus, D., Tran, D. K., Hayden, L., Carah, N., Luong, K., & Newton, G. (2026). The mobile observation of advertising toolkit: A tool for understanding ephemeral and sequenced social media data. *Computational Communication Research*, *8*(1), 1–15. https://doi.org/10.5117/CCR2026.1.9.OBEI

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2023). Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, *18*(2), 124–141. https://doi.org/10.1080/19312458.2023.2181319

Papacharissi, Z. (2023). What is privacy. In S. Trepte & P. Masur (Eds.), *Routledge handbook of privacy and social media* (pp. x–xiii). Routledge.

Perriam, J., Birkbak, A., & Freeman, A. (2020). Digital methods in a post-API environment. *International Journal of Social Research Methodology*, *23*(3), 277–290. https://doi.org/10.1080/13645579.2019.1682840

Pfiffner, N., Witlox, P., & Friemel, T. N. (2024). Data donation module: A web application for collecting and enriching data donations. *Computational Communication Research*, *6*(2), Article 4. https://doi.org/10.5117/CCR2024.2.4.PFIF

Poell, T., Nieborg, D., & van Dijck, J. (2019). Platformisation. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1425

Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, *9*(4), 1–28. https://doi.org/10.14763/2020.4.1535

Rogers, R. (2017). Digital methods for cross-platform analysis. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The Sage handbook of social media* (pp. 91–110). Sage. https://doi.org/10.4135/9781473984066.n6

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064. https://doi.org/10.1126/science.346.6213.1063

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Auditing algorithms: Research methods for detecting discrimination on internet platforms* [Paper presentation]. 64th Annual Meeting of the International Communication Association, Seattle, WA, United States. https://ai.equineteurope.org/system/files/2022-02/ICA2014-Sandvig.pdf

Schoch, D., Chan, C. H., Wagner, C., & Bleier, A. (2024). Computational reproducibility in computational social science. *EPJ Data Science*, *13*(1), Article 75. https://doi.org/10.1140/epjds/s13688-024-00514-w

Selling, L. K., Iglesias Keller, C., Ohme, J., Klinger, U., & de Vreese, C. (2025). *Data access for researchers under the digital services act: From policy to practice* (Weizenbaum Policy Paper No. 14). Weizenbaum Institut. https://hdl.handle.net/10419/326736

Sinnott, R. O., Li, Q., Mohammad, A., & Morandini, L. (2023). The Australian Digital Observatory: Social media collection, discovery and analytics at scale. In C.-H. Hsu, M. Xu, H. Cao, H. Baghban, & A. B. M. Shawkat Ali (Eds.), *Big data intelligence and computing* (pp. 316–327). Springer. https://doi.org/10.1007/978-981-99-2233-8_23

Teepapal, T. (2025). AI-driven personalization: Unraveling consumer perceptions in social media engagement. *Computers in Human Behavior*, *165*, Article 108549. https://doi.org/10.1016/j.chb.2024.108549

Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media + Society*, *7*(1). https://doi.org/10.1177/2056305121988929

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 505–514. https://doi.org/10.1609/icwsm.v8i1.14517

Turk, K., Pastrana, S., & Collier, B. (2020). A tight scrape: Methodological approaches to cybercrime research data collection in adversarial environments. In A. Armando, G. Pellegrino, & P. Prinetto (Eds.), *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (pp. 428–437). IEEE. https://doi.org/10.1109/EuroSPW51379.2020.00064

van Atteveldt, W., Trilling, D., & Calderon, C. A. (2022). *Computational analysis of communication*. Wiley.

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine-learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Wähner, M., Deubel, A., Breuer, J., & Weller, K. (2024). "Don't research us"—How Mastodon instance rules connect to research ethics. *Publizistik*, *69*, 357–380. https://doi.org/10.1007/s11616-024-00855-6

About the Authors

Bogdan Mamaev is the research training manager at the Australian Internet Observatory (AIO). A political science researcher by training, his work lies in applying computational tools to social science research and automating data collection to create protest event datasets.

Daniel Angus is professor of digital communication at the Queensland University of Technology and director of the Digital Media Research Centre. He is also a chief investigator in the ARC Centre of Excellence for Automated Decision-Making and Society and leads the QUT node of the Australian Internet Observatory. Daniel's research examines the dynamics of digital platforms, with a particular focus on algorithmic systems, political communication, and the observability of online environments.

Amanda Lawrence has a PhD in media and communication from RMIT University. She is a senior advisor at the Australian and New Zealand School of Government (ANZSOG). Prior to this she was engagement lead and program director at the Australian Internet Observatory.

Richard Sinnott (BSc, MSc, PhD) is professor of applied computing systems at the University of Melbourne. He has been technical lead on a multitude of large-scale international projects with emphasis on big data and security worth over $500m. This includes numerous projects in the defence, intelligence, and biomedical domains. He has over 500 peer-reviewed publications across a range of computing and application-specific domains.