

Political Communication Report
Summer 2026 - Issue 33
“Political Communication Research under Data Constraints”

CanberraInbox: Building a New Data Infrastructure for Political Communication¹

Rosa Rosmery Soto Ruidias, Australian National University

Daniel Casey, Australian Catholic University

Political behavior is increasingly mediated, digitally traceable, and both more observable and less observable at the same time. Researchers face mounting constraints in accessing the data required to study political communication systematically. Platform restrictions, application programming interface (API) shutdowns, legal limitations, and proprietary data regimes have transformed what was once a relatively open empirical environment into a fragmented and unequal landscape. At the same time, concerns about bias, misinformation, and disinformation have further heightened the importance of studying public and political communication and messaging from a broad range of sources (Broda & Strömbäck, 2024; Governance Responses to Disinformation, 2020). In response to these changing dynamics, scholars must move beyond reliance on existing datasets and invest in independent, sustainable data infrastructure. CanberraInbox² illustrates this approach by systematically capturing forms of political communication that are often overlooked in existing research.

The “age of restricted data,” that this edition of Political Communication Report is structured around, is not merely as a technical constraint, but a structural turning point for the field. It will require scholars to consider less observable data sources and create new ongoing data collection and processing to study political communication. Rather than treating restricted access solely as a limitation, it can be approached as an opportunity to address long-standing empirical and normative biases, particularly the overreliance on platform-mediated communication and the relative neglect of alternative data infrastructures.

From Platform Dependence to Data Construction

Over the past decade, computational political communication research has been heavily shaped by access to large-scale platform data, particularly from Twitter/X, Facebook, and other social media environments (Blakey, 2024; Hasrullah & Suherman, 2025; Trezza, 2023). Studies demonstrated the analytical power of such data, enabling researchers to map networks, measure polarization, and track elite discourse at scale (Ohme et al., 2024). However, this context has been always contingent on a restricted infrastructure. Scholars have been dependent on

¹ Copyright © 2026. Soto Ruidias, R., R. & Casey, D. Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at <http://politicalcommunication.org>

² <https://enewsletters.shinyapps.io/canberrainbox/>

privately owned platforms to provide access according to their own terms and timeline (Couldry & Gao, 2026).

That infrastructure has now eroded. API restrictions, increased costs, and data access limitations have made it significantly more difficult (Baumgartner et al., 2020; Bruns, 2021; Chen et al., 2024), if not impossible, for many researchers to replicate earlier studies or build comparable datasets. This has produced a form of platform dependency, where research agendas may shaped not only by theoretical priorities, but also by what data happen to be available (Silverman, Brandon, 2025). In effect, restricted data risks reinforcing existing inequalities within the field while narrowing the empirical scope of political communication research. For example, the increased costs associated with API access are likely to exacerbate differences between researchers with large research budgets and those without. At the same time, there is an ongoing risk associated with the ‘streetlight effect’ or the ‘drunkard’s search’ (Bimber, 2015) if we limit our research to those datasets where we have easy, cheap access. For example, CanberraInbox only captures e-newsletters, with an unknown circulation. We miss out entirely on physical newsletters, because they are much harder to collect and store (Koop & Marland, 2013).

Seeing the “Invisible” Layer of Representation

The CanberraInbox project emerged not only from these broader challenges, but also from a practical insight grounded in firsthand experience. One of this essay’s authors and dataset creator (Dr Casey) previously worked in a parliamentary office, where one of his professional responsibilities was drafting physical newsletters sent to constituents. This experience revealed a dimension of political communication largely invisible to both researchers and the public: the routine, strategic, and carefully curated communication through which legislators engage directly with their constituents. This form of communication sits largely outside media coverage, social media visibility, and most traditional datasets, yet it is central to how representation is practiced, as it is within these newsletters that legislators explain their work, claim credit, signal priorities, and construct relationships with voters.

CanberraInbox was designed to make this “hidden” layer observable. By subscribing to the newsletter distribution lists of all Members of Parliament (MPs) and systematically collecting their emails, the dataset captures a communication channel that is direct, intentional, and theoretically meaningful. Importantly, this also creates a kind of “bird’s-eye view” of representation that is typically unavailable (Casey, 2025b). It allows researchers to observe what representatives say to their subscribers, which may differ from their public-facing communication. In a context of widespread but often poorly informed political work, this has broader democratic relevance. These e-newsletters are usually taxpayer funded, yet national archives and congressional/parliamentary libraries do not adequately record or maintain access to these important official documents. When DCInbox was established, Cormack (2017: 27) reported that “the Library of Congress reportedly stores hard copy versions of each of these electronic communications,” but did not make electronic versions available. Initial investigations with the Australian Parliamentary Library indicated that they did not maintain any records of MPs physical or e-newsletters.

Many different political communication research questions could be addressed using these data, often in conjunction with other data sources. DCInbox has been used for dozens of different projects, including exploring agenda-setting and framing; differences between parties in communication styles; and emphasis across specific policy areas (e.g. Casey, 2025a).

Designing the Dataset: Minimalism and Extensibility

The development of CanberraInbox was informed by existing projects such as DCInbox³ and UKMPInbox⁴, as well as collaboration with a computer science research assistant (Ms Soto). This reflects the importance of a multidisciplinary approach, combining political science and data science. While political science provides the theoretical framework and defines the relevant constructs, data science contributes the technical expertise required to build scalable, structured, and reusable datasets.

Rather than attempting to include all possible metadata at the point of collection, the dataset focuses on a small set of core variables: full-text newsletters, timestamps and a stable MP identifier (a unique identifier, created by the parliamentary library [PHID]; name; chamber; electorate; party; parliamentary term). The rationale is that most additional variables (e.g., electorate characteristics, or electoral margins) can be merged later by individual researchers. What matters at the point of data construction is ensuring a stable, consistent, and extensible core structure (Jonker, 2025). This design also facilitates longitudinal analyses; for example, researchers can examine changes in an MP's political communication over time by linking this dataset to other that captures parliamentary events, party switches, ministerial promotions, or other developments occurring during the same period. For those variables that are likely to change during a parliamentary term (e.g. a representative changes party, or gets a promotion), it was sometimes difficult to ensure that the data was updated quickly and automatically. Where possible, we would encourage other dataset creators to either focus on a small set of core variables, or leverage other datasets/packages that you are confident will be regularly updated. In this regard, we relied on the AusPH package (Leslie, 2024).

This design choice proved particularly important given the longitudinal and evolving nature of the dataset. By anchoring the data around a unique identifier of member of the parliament, CanberraInbox is designed to remain interoperable with external datasets and adaptable to future research needs. More broadly, this reflects a key lesson: good dataset design is not about maximizing variables at the point of collection, but about ensuring long-term flexibility and reusability (Koesten et al., 2020).

³ <https://www.dcinbox.com>. Separately, Dr Adam Ozer started UK MP Inbox (https://ukmpinbox.shinyapps.io/uk_mp_inbox), which means that there is now similar comparable data across Australia, the UK and the USA.

⁴ http://ukmpinbox.shinyapps.io/uk_mp_inbox

From Dataset to Infrastructure

CanberraInbox is not only a dataset; it is a simple structure ongoing data infrastructure (See Figure 1). Maintaining it involves a continuous workflow that integrates data collection, processing, validation, and dissemination. This includes:

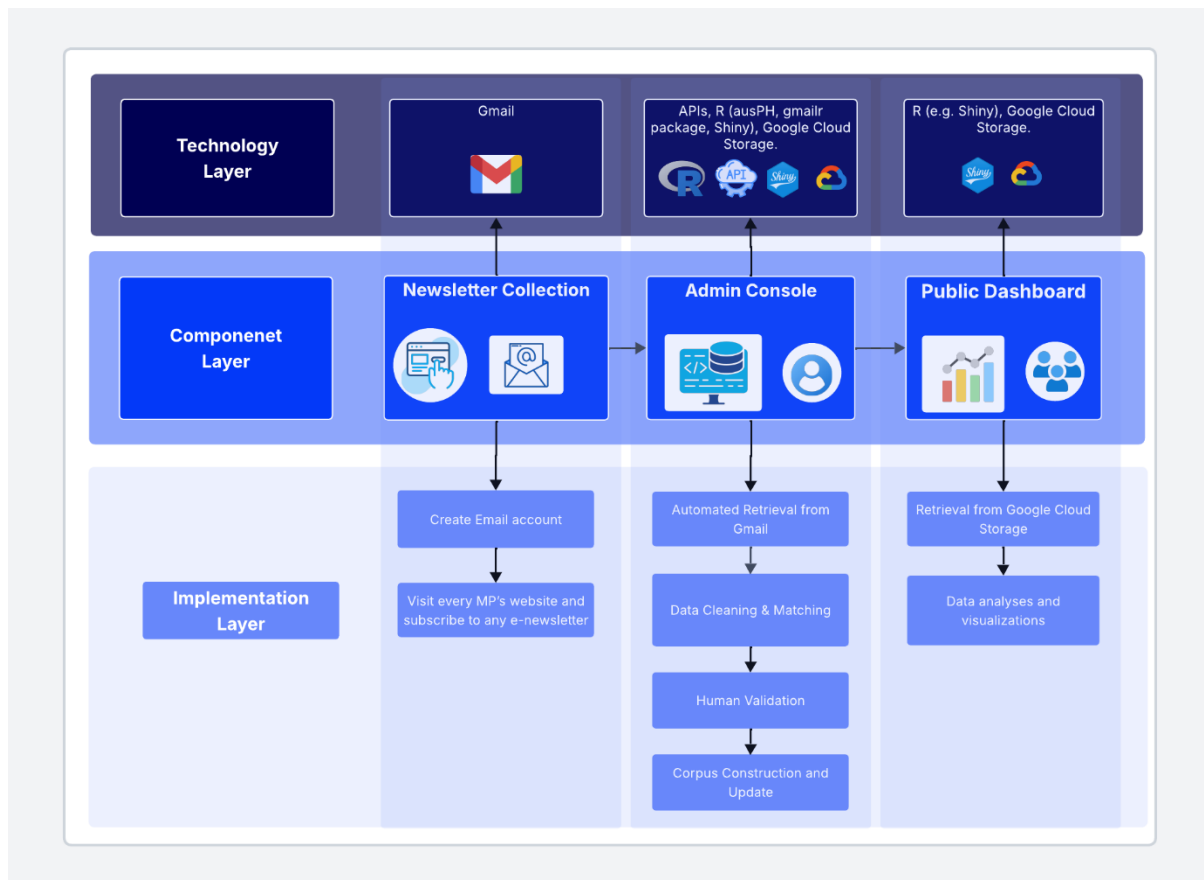
- continuous collection of newsletters through systematic subscriptions
- automated preprocessing and sender attribution
- structured storage and archiving
- regular updates as new newsletters are received

To support this process, an administrative interface was developed to manage the pipeline. This interface allows for the retrieval of new data, facilitates human validation of automated matching, and enables the controlled updating of the dataset and associated corpus. Importantly, this introduces a human-in-the-loop validation stage, where automated processes are reviewed and corrected before being incorporated into the final dataset.

In parallel, a public-facing interface provides access to the dataset for researchers, journalists and members of the public. This interface allows users to search, filter, visualize, and download the data, lowering the technical barriers to use and enabling both academic and non-academic engagement. Together, these components illustrate that CanberraInbox operates as a multi-layered system, combining automated processing, human curation, and public dissemination.

This highlights a critical but often overlooked point: the value of datasets lies not only in their initial construction, but in the ongoing labor required to sustain them. In contexts such as Australia, where funding and institutional support for such work are limited, this labor, spanning maintenance, troubleshooting, and incremental improvement, remains significantly under-recognized. The CanberraInbox architecture and workflow may also serve as a model for implementing other systems that require similar methods.

Figure 1: CanberraInbox Architecture & Workflow



Methodological Challenges and Open Questions

Building an independent dataset does not eliminate methodological challenges. One of the attractions of these e-newsletters was the potential that these were largely under-studied and under-examined by both other researchers and the media, and thus maybe avoids the traditional Schrodinger's effect / observation problem in research. Once legislators become aware that their communications are being systematically collected and analyzed, they may alter their behavior. Whether this constitutes a source of bias or simply reflects the inherently performative nature of political communication remains an open question. However, this is a standard issue in political communications research and may itself be a meaningful object of analysis.

Second, there are ambiguities in the unit of analysis. In more party-centric systems, such as New Zealand, newsletters may be distributed centrally rather than by individual MPs, complicating attribution. Similarly, we have been informally told that some MPs send multiple targeted newsletters (e.g., segmented by postcode), raising questions about whether any single subscription captures the full scope of their communication. However, it would be difficult to systematically identify this.

Third, there are ongoing technical and administrative challenges, including:

- matching newsletters sent from multiple email addresses to a single MP identifier
- filtering out irrelevant or misdirected communications
- maintaining coverage as MPs enter, exit, or change roles within parliament

These challenges underscore that dataset construction is not a one-off task, but an iterative process requiring continuous refinement and oversight. This perspective is also supported by experts and should be reflected in the ongoing improvement of the development process (Venkatasubramanian et al., 2024).

Expanding the Empirical Scope

There are a range of ways that future researchers could expand the scope of this work. We initially started collecting similar e-newsletters from the other two remaining Anglosphere countries, New Zealand and Canada. These were discontinued, however, due to some of the methodological challenges outlined above, primarily the lack of local knowledge hampered our ability to maintain coverage.

While the current dataset focuses on textual content, the underlying data structure allows for future extensions. Because newsletters are stored in HTML format, it is technically feasible to archive images and other embedded elements. Incorporating such data would enable the study of visual political communication, though it would require additional storage capacity and processing resources.

More broadly, CanberraInbox illustrates how datasets can evolve incrementally over time. By designing for extensibility from the outset, new layers of data can be added without compromising the integrity of the original structure. This reinforces the importance of viewing datasets as dynamic and evolving resources rather than static products, while adopting a flexible and simple design that balances information needs with available resources.

If others are interested in starting something similar, we are happy to share our experiences. The initial set-up is quick and easy – a dedicated Gmail account, and the time to visit every MP’s website and subscribe to any e-newsletter. There were zero upfront costs. The R script to access the Gmail API and download the emails to a CSV file are similarly relatively simple. Code is available on GitHub⁵

Rethinking Data in Political Communication

The broader implication is that the age of restricted data should push the field toward a different model of empirical research. Rather than relying on externally controlled data sources, researchers can: build datasets aligned with theoretical constructs, make data generation processes transparent, and treat absence and variation as analytically meaningful

⁵ <https://github.com/research-projects-info/canberraInbox>

This shift also requires rethinking how the field values research contributions. The construction and maintenance of datasets, tools, and infrastructures should be recognized as central scholarly work, rather than peripheral technical support.

Conclusion

Restricted data has exposed the fragility of a research paradigm built on platform access. But it has also created an opportunity to rethink the empirical foundations of political communication research. Projects like CanberraInbox demonstrate that it is possible to build independent, theoretically grounded datasets that expand the scope of the field. More importantly, they show that doing so requires not only careful technical design, but sustained infrastructural work, ongoing maintenance, and methodological reflexivity. However, this does not need to be overly complex; in fact, it can be achieved through the adoption of a simple architecture.

References

- Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 830–839. <https://doi.org/10.1609/icwsm.v14i1.7347>
- Blakey, E. (2024). The Day Data Transparency Died: How Twitter/X Cut Off Access for Social Research. *Contexts*, 23(2), 30–35. <https://doi.org/10.1177/15365042241252125>
- Bimber, B. (2015). What's next? Three challenges for the future of political communication research. *New technologies and civic engagement*, 215–233.
- Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: Lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, 48(2), 139–166. <https://doi.org/10.1080/23808985.2024.2323736>
- Bruns, A. (2021). After the ‘APIcalypse’: Social media platforms and their fight against critical scholarly research. *Disinformation and Data Lockdown on Social Platforms*, 14–36.
- Casey, D. (2025a). *Australia’s Foreign Policy in the Trump Era: Balancing Responsiveness and Responsibility?*
- Casey, D. (2025b). *CanberraInbox: Political Communication, the Personal Vote and Representation Styles—Studying Legislators’ e-Newsletters in Australia*. *Legislative Studies Quarterly*, 50(3), e70004. <https://doi.org/10.1111/lsq.70004>
- Chen, Y., Sherren, K., Lee, K. Y., McCay-Peet, L., Xue, S., & Smit, M. (2024). From theory to practice: Insights and hurdles in collecting social media data for social science research. *Frontiers in Big Data*, 7, 1379921. <https://doi.org/10.3389/fdata.2024.1379921>
- Couldry, N., & Gao, G. (2026). The Space of the World, Data Colonialism, and Social Contract: Reconstructing Digital Futures—A Dialogue with Nick Couldry. *Communication and the Public*, 11(1), 46–52. <https://doi.org/10.1177/20570473261424566>
- Governance responses to disinformation: How open government principles can inform policy options (OECD Working Papers on Public Governance No. 39; OECD Working Papers on Public Governance, Vol. 39). (2020). <https://doi.org/10.1787/d6237c85-en>
- Hasrullah, H., & Suherman, A. (2025). From Speeches to Tweets: The Mapping of Trend and Evolution of Political Communication in Digital Media. *Thammasat Review*, 28(2), 205–231.
- Jonker, T. K., Alexandra. (2025, July 25). *What Is a Data Architecture?* | IBM. <https://www.ibm.com/think/topics/data-architecture>
- Koesten, L., Vougiouklis, P., Simperl, E., & Groth, P. (2020). Dataset Reuse: Toward Translating Principles to Practice. *Patterns*, 1(8), 100136. <https://doi.org/10.1016/j.patter.2020.100136>
- Koop, R. and Marland, A. (2012), Insiders and Outsiders: Presentation of Self on Canadian Parliamentary Websites and Newsletters. *POI*, 4: 112–135. <https://doi.org/10.1002/poi3.13>
- Leslie, Pat. 2024. “AusPH: An R Package to Retrieve Data from the Parliamentary Handbook of the Commonwealth of Australia.” <https://github.com/palesl/ausPH>
- Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2024). 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>
- Silverman, Brandon. (2025). *Why Commercial Tools Can Scrape Social Media But Researchers Can't* | TechPolicy.Press. Tech Policy Press.
<https://www.techpolicy.press/why-commercial-tools-can-scrape-social-media-but-researchers-cant/>
- Trezza, D. (2023). To scrape or not to scrape, this is dilemma. The post-API scenario and implications on digital research. *Frontiers in Sociology*, 8, 1145038.
<https://doi.org/10.3389/fsoc.2023.1145038>
- Venkatasubramanian, H., Anisa, M., Ramesh, S. R., Subbiah, B., & Vijayaraghavan, A. (2024). Application of Data Analytics in IT project Management: Improving efficiency, Risk Mitigation. *Proceedings of the 6th International Conference on Information Management & Machine Intelligence*, 1–9.
<https://doi.org/10.1145/3745812.3745882>



[Rosa R. Soto Ruidias](#) is a PhD student at the ANU School of Computing. She was the research assistant and coder on the CanberraInbox project. Her research interests include Data Fairness and AI for Public Good, with a current focus on investigating political biases in LLMs and their societal implications. Her work is built on seven years of experience across Peru and Australia, where she contributed with IT implementations—including web platforms, datasets, and analytical dashboards—alongside quality management and educational accreditation. By leveraging learning analytics and data science, she aims to develop equitable AI systems that enhance public service and educational transparency.



[Dr. Daniel Casey](#) is a lecturer in politics and international relations at the Australian Catholic University. He completed his PhD in August 2024 at ANU, examining letters from members of the public to Australian Prime Minister Howard - who writes; why they write; and the impact of the letters on public policy and the political agenda. His broader research interests include elite-mass linkages, with a focus on different forms of communication between the public and leaders; public policy and public administration; and representation. These research interests are driven by his 15-year public service career, including working for members of parliament.