IFLA WLIC 2026 BUSAN
LIBRARIES POWERING TRANSFORMATION
World Library and Information Congress
90th IFLA General Conference
10–13 August 2026, South Korea

Submitted on: June 29, 2026

Powering Trust: Tracking and Preserving Government Data References in the News Media

Groenendyk, Michael
Concordia University Library, Montreal, Canada
E-mail address: Michael.Groenendyk@concordia.ca

Kung, Janice Y.
University of Alberta Libraries, Edmonton, Canada
E-mail address: janice.kung@ualberta.ca

Abstract:

Journalists frequently cite government data, yet poor citation practices and post-publication dataset modifications hinder independent verification (Camaj et al., 2025; Goes, 2025). This leaves data citations highly vulnerable to AI hallucination and misattribution.

Our paper analyzes 3,290 in-text citations from Canadian news media using a fine-tuned BERT model and RAG-powered AI agents to enrich data citations. To ensure integrity, we stored open data metadata and citations as a preservation layer on the Arweave blockchain, creating permanent, tamper-evident records of dataset versions via hashing.

Our findings reveal that news media citations typically lack crucial metadata, such as versions and URLs, needed for verification. Crucially, we demonstrate how large language models (LLMs) can track citations across media, and how blockchain infrastructure provides the verifiability lacking in current government portals. While focused on the Canadian context, this project offers an internationally transferable model for data accountability. Librarians are uniquely positioned to build this decentralized infrastructure. Serving as a critical verification layer for humans and LLMs, these workflows ensure digital collections remain trustworthy in a rapidly evolving technological landscape.

Keywords: data citation; open data verification; blockchain preservation; artificial intelligence; digital journalism

Introduction

Government data has become an increasingly important information source for journalists, providing critical insight into economic, social, and political issues and helping shape public discourse around government policy (Camaj et al., 2025). In Canada, the Government of Canada launched its Open Data Portal (data.gc.ca) in March 2011 and has since expanded it to include over 4,000 datasets, and over 10,000 when geospatial layers are included (Treasury Board of Canada Secretariat, 2020).

Within news media, however, government data is frequently cited informally, without clear authorship attribution, consistent naming conventions, or persistent identifiers (Park et al., 2018). This ambiguity makes measuring the impact of open data through conventional bibliometric approaches difficult (Buneman et al., 2020). Compounding this problem, government data is not static (Goes, 2025). Datasets referred to in news articles may later be revised, updated, or restructured without notice, meaning the evidence base underlying a published article can change after publication. This is especially problematic in the context of large language models (LLMs), which are increasingly used to summarize and generate content referencing government statistics and are vulnerable to misattributing or fabricating citations (Huang & Chang, 2023). Without a stable, verifiable record of what a dataset contained at the time it was cited, AI-generated research risks presenting outdated or modified data as current.

To address both unclear citations and post-publication dataset modification, this project employs three complementary methods to identify, verify, and preserve citations of Government of Canada data in Canadian news media: locally run fine-tuned BERT models for citation detection; retrieval-augmented generation (RAG) AI agents for citation enrichment and dataset matching; and a preservation and integrity verification layer built on the Arweave blockchain.

This research contributes to broader discussions on data citation standardization, open data impact measurement, and the role of government data in informing public debate. It also proposes a practical infrastructure model that librarians are well positioned to build and maintain. As governments worldwide invest in open data programs, robust frameworks for citation tracking, integrity verification, and preservation become essential for accountable journalism, for reproducible research, and for ensuring AI systems have access to stable, verifiable ground truth (Groenendyk, 2023).

Literature Review

Data Citation Detection

Detecting citations of datasets has drawn increasing attention.The Make Data Count initiative has worked to standardize data impact metrics (Kratz & Strasser, 2015), collaborating with Counter to create a Code of Practice for Research Data Usage Metrics (Fenner et al., 2018), and more recently partnering with the Chan Zuckerberg Initiative and the Welcome Trust to build an open global Data Citation

Corpus (Vierkant, 2023). In 2021, the Coleridge Initiative launched a competition challenging data scientists to use natural language processing to demonstrate open data impact (Coleridge Initiative, 2021). DataCite has advanced this work by implementing persistent identifiers for datasets (DataCite, 2023), while CrossRef has integrated these identifiers to illustrate connections between open data and research outputs (CrossRef, 2023).

Commercial platforms have also entered this space, with the Web of Science Data Citation Index providing a dedicated tool for tracking dataset citations within scholarly research (Robinson-García et al., 2016).While studies have explored how LLMs can track references to traditional publications (Huang & Chang, 2023; Zhang et al., 2023), few have investigated their potential for tracking references to datasets, where citation patterns are highly inconsistent. A notable exception is work by the Chan Zuckerberg Initiative on tracking software references in academic literature, which provides a promising methodological foundation (Istrate et al., 2022/2024).

Data Integrity and the Limits of Current Infrastructure

While progress has been made on standardizing data citations, little research has examined how often government data is modified post-publication. Most citation frameworks assume a cited dataset remains accessible and unchanged after publication. In practice, government datasets can be revised, updated, or removed without notice. Existing open data portals typically lack versioning systems that would allow users to retrieve the specific version cited in a news article (Goes, 2025).

This is particularly problematic for government data, which Buneman et al. (2020) note is already the most inconsistently cited of any data type. When the underlying data is also subject to silent modification, citations are vague, the resources they point to are unstable, and independent verification of data-driven claims becomes extremely difficult.

Persistent identifiers like DOIs address part of this problem by ensuring a citation resolves to a stable location, but they do not guarantee the content at that location is unchanged. A DOI can resolve to a dataset substantially revised since the citing work was published, with no mechanism for readers to verify the version of record. This gap between identifier persistence and content persistence is a significant unresolved challenge.

AI Hallucinations and the Need for Verifiable Ground Truth

The emergence of LLMs has further complicated citation integrity. Researchers have documented how LLMs are prone to generating plausible-sounding but fabricated citations, a phenomenon known as hallucination (Huang & Chang, 2023). Formulaic phrases like "according to Statistics Canada" dominate media coverage but provide LLMs with organizational attribution and no dataset-level specificity. Without a stable, verifiable record of what a given government dataset contained at a specific point in time, it is impossible to determine whether an LLM-generated claim about government data is accurate, outdated, or fabricated.

Methodology

The methodology has three core components: citation identification using a fine-tuned BERT model, citation enrichment using RAG-powered AI agents, and metadata preservation on the Arweave blockchain.

Corpus Compilation

To obtain news article text citing Government of Canada data, the websites of Canadian news publishers (such as the CBC) were searched via Google for articles mentioning "data" alongside a Government of Canada organization name. Identified articles were reviewed by human researchers, and confirmed data-citing sentences were saved to a CSV file for BERT fine-tuning. In total, 3,290 data-citing sentences were identified within 12,883 news articles. Training data, models, and code are openly available on the project's GitHub and HuggingFace pages.

Citation Detection

Candidate citation sentences were identified using a fine-tuned BERT classification model, run locally on a laptop. Sentences scoring above a probability threshold were flagged as candidate citations. Additional filters were applied to remove boilerplate legal text (e.g., copyright notices) and to require at least one government-domain signal (such as an agency name, acronym, or domain term like "census"). Running the model locally meant the only cost was electricity, an important consideration for scalability.

.

LLM Enrichment

Sentences passing through the filters were sent to an LLM enrichment agent built on the Amazon Bedrock Converse API. The pipeline was run with two models in parallel (Claude Sonnet 4.6 and Claude Haiku 4.5) to compare enrichment quality and cost.

Each enrichment call assembled four contextual layers: the article title, abstract, the surrounding paragraph, and the citation sentence itself. The model was instructed to act as a research librarian and given access to a search tool that queried the Government of Canada Open Data Portal. The model iteratively issued search queries until it found a satisfactory match or exhausted its attempts, then returned a structured JSON record.

Temporal Stability Check

For each citation where both an article publication date and a dataset modification date were available, the dates were compared. Citations matched to datasets modified *after* the article's publication were flagged as potentially unstable, indicating that the dataset as it currently exists may differ from the version available at the time of writing.

Arweave Metadata Preservation

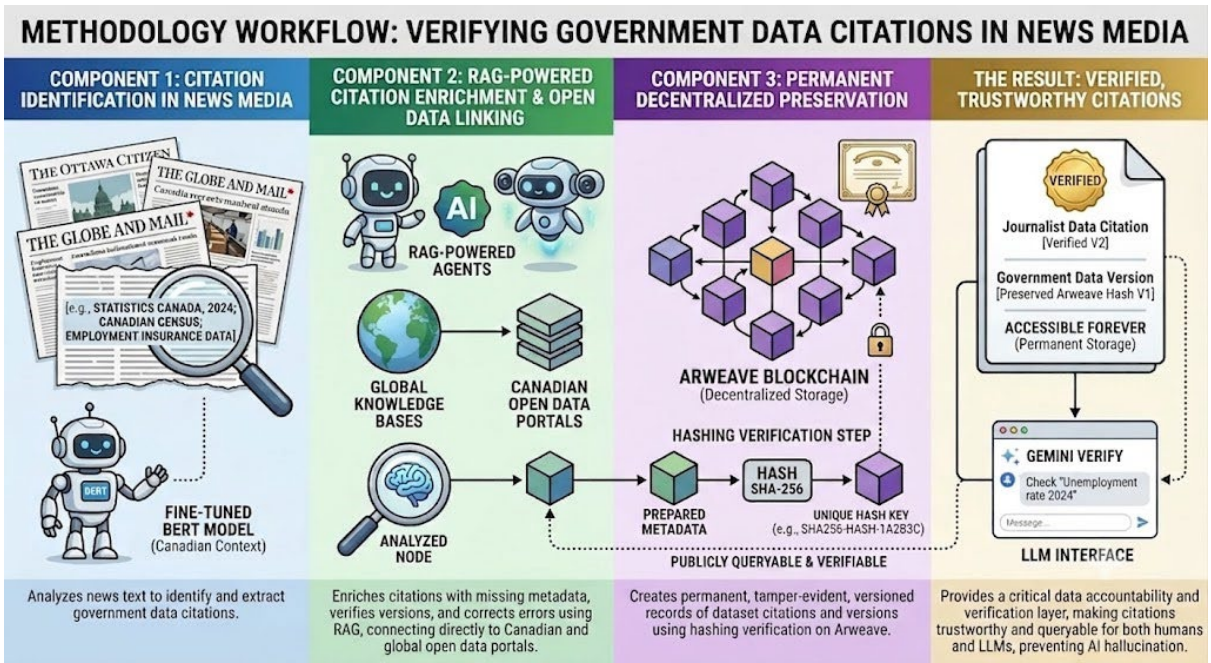
A core component of this project is a preservation layer built on Arweave, a decentralized storage protocol designed for permanent data preservation. Records stored on Arweave are cryptographically tamper-evident and distributed across a peer-to-peer network, making retroactive modification effectively impossible and ensuring long-term accessibility without dependence on any single institution.
For each citation, two categories of records were stored on Arweave:

1. Citation and metadata records: The citation sentence, news outlet, publication date, LLM enrichment metadata, and the Dublin Core metadata for the matched dataset (retrieved from open.canada.ca at the time of processing).

2. Integrity hashes: To provide cryptographic evidence of dataset integrity *without* storing dataset content, the pipeline streamed each dataset resource and computed a SHA-256 hash of its byte content in-flight. Only the hash digest was stored; the underlying content was discarded. By re-hashing the same resource at any future point and comparing it to the on-chain record, any researcher can independently verify whether the dataset has been modified since the citation was processed. A change in the digest is evidence of modification; an identical digest confirms byte-for-byte integrity.

These on-chain records provide a stable reference point for researchers and journalists, a verifiable corpus of ground-truth data citations for AI systems, and a model of decentralized preservation infrastructure for librarians and archivists.

Figure 1: Methodology Workflow Architecture Diagram



Limitations

The search strategy relied on articles using both a government organization name and the word "data," likely missing articles that cite government data without using that specific term. The human review process used to construct training data also involved interpretive judgment that may have introduced inconsistencies. Finally, because the hashing approach stores only the digest (not the content), a hash mismatch will correctly signal modification, but the original content cannot be recovered from the Arweave record alone, a deliberate trade-off between storage cost and verifiability.

Analysis

Citation Detection Performance

The fine-tuned BERT model correctly identified 3,182 of the 3,290 human-confirmed citing sentences (96.7%), but in a fraction of the time required for human review.

Of the BERT-identified citing sentences, 93.5% were successfully enriched with full metadata by the Haiku-powered agent: 47.2% with high confidence (a clear match to a specific dataset), 37.6% with medium confidence (a plausible but ambiguous match), and 9% with low confidence (a tentative match only).

The Sonnet 4.6 agent performed slightly better, enriching 93.8% of citations with a similar confidence distribution but at three times the cost. Haiku enriched the full corpus for $16.52 USD ($0.0052 per citation), while Sonnet cost $51.23 ($0.0161 per citation). The 1.1% improvement in enrichment rate did not justify the cost difference for this use case. The BERT models, run locally on a laptop, were essentially free.

Table 1: Enrichment Confidence by Model

Confidence	Haiku 4.5 (N / %)	Sonnet 4.6 (N / %)
High	1,503 / 47.2%	1,576 / 49.5%
Medium	1,196 / 37.6%	1,124 / 34.7%
Low	286 / 9.0%	320 / 10.6%
Not Enriched	197 / 6.2%	162 / 5.1%
Total	**3,182 / 100%**	**3,182 / 100%**

Top Cited Organizations

Statistics Canada emerged as the clear leader, cited in 1,593 sentences, roughly half of all citations, and far more than the second-cited Royal Canadian Mounted Police (709) and Health Canada (522). Other notable sources included the Canada

Revenue Agency, the Canada Border Services Agency, and the Canadian Security Intelligence Service. Many sentences cited multiple organizations together (e.g., "data from Statistics Canada and Health Canada"). Cited Statistics Canada datasets covered economic indicators, employment, population change, and social dynamics, while other agencies' datasets appeared most often in health and crime reporting.

Citation Verifiability

The majority of government citations in news media provided only an attribution to the government department, often through phrases like "according to Statistics Canada" or "data from the Canada Revenue Agency," without specifying a dataset name, URL, publication date, or version identifier. Only about 32% of citations included sufficient detail to identify a specific dataset, and even among these, very few included a direct URL or version identifier.

Table 2: Citation Verifiability

Verifiability	N	%
Verifiable	1,018	32.0%
Partially verifiable	1,027	32.3%
Not verifiable	940	29.5%
Not enriched	197	6.2%
Total	**3,182**	**100%**

This metadata gap has direct implications for both human verifiability and AI reliability. For human readers, the absence of dataset-level specificity makes independent verification of data-driven claims extremely difficult. For AI systems trained on news media text, these vague attributions provide insufficient structured information to reconstruct or verify underlying data, creating conditions conducive to hallucination (Huang & Chang, 2023).

Conclusion

The findings of this study point toward a clear institutional gap: the infrastructure needed to make government open data citations verifiable does not currently exist within government portals, journalism workflows, or academic publishing systems. Government portals lack robust versioning. Journalism organizations lack both the technical capacity and institutional mandate to archive the datasets that inform their reporting, and the industry lacks standardized citation practices—relying instead on ephemeral hyperlinks or casual in-text attributions. Academic citation standards, while more rigorous, still do not systematically address dataset content stability over time.

Librarians are uniquely positioned to fill this gap. The professional expertise of library and information science encompasses both the technical dimensions of digital preservation and the institutional relationships needed to build infrastructure that serves multiple stakeholder communities. Libraries have a long history of maintaining trusted, neutral preservation infrastructure for materials that governments and publishers do not preserve themselves, and the challenge of open data citation verifiability is a natural extension of this mandate.

The blockchain preservation model developed in this project offers a concrete, internationally transferable template. The core workflow of identifying citations using machine learning, retrieving and storing dataset metadata, and creating tamper-evident records on a decentralized network, is not specific to Canada. Libraries in other countries could adapt this model to their national open data ecosystems, creating a distributed global network of preservation nodes that collectively provide the verifiability infrastructure no single government currently offers.

This infrastructure serves a dual purpose. For human researchers, journalists, and citizens, it provides a reliable mechanism for verifying data-driven claims. For AI systems, it provides a structured, tamper-evident corpus that can serve as ground truth, reducing hallucination and misattribution risks (Huang & Chang, 2023).
The international dimensions of this challenge are worth emphasizing for an IFLA audience. The problems identified here, such as inconsistent citation practices, unstable government datasets, and AI systems prone to misattributing government statistics, are not unique to Canada. They reflect structural features of how government data is published, cited, and consumed across jurisdictions worldwide. IFLA is well positioned to facilitate the cross-jurisdictional coordination needed to develop shared standards for open data citation preservation, and to advocate for libraries as essential infrastructure providers in the emerging ecosystem of AI-assisted research and journalism.

As governments continue to prioritize open data, and as AI becomes more deeply integrated into research and public discourse, the need for robust, decentralized verification infrastructure will only grow. Libraries that pioneer these workflows today will be well positioned to serve as the trusted verification layer that this emerging ecosystem requires.

Acknowledgments

This research was made possible through a Social Sciences and Humanities Research Council (SSHRC) Insight Development Grant.

References

Buneman, P., Christie, G., Davies, J. A., Dimitrellou, R., Harding, S. D., Pawson, A. J., Sharman, J. L., & Wu, Y. (2020). Why data citation isn't working, and what to do about it. *Database, 2020*. https://doi.org/10.1093/databa/baaa022

Camaj, L., Martin, J., & Lanosga, G. (2025). The impact of public transparency infrastructure on data journalism: A comparative analysis between information-rich and information-poor countries. *Digital Journalism, 13*(2), 268–287. https://doi.org/10.1080/21670811.2022.2077786

Coleridge Initiative. (2021). *Coleridge Initiative—Show US the data*. https://kaggle.com/competitions/coleridgeinitiative-show-us-the-data

CrossRef. (2023). *Data and software citation deposit guide*. https://www.crossref.org/documentation/reference-linking/data-and-software-citation-deposit-guide/

DataCite. (2023). *What we do*. https://datacite.org/what-we-do/

Fenner, M., Lowenberg, D., Jones, M., Needham, P., Vieglais, D., Abrams, S., Cruse, P., & Chodacki, J. (2018). *Code of practice for research data usage metrics release 1* [Preprint]. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.26505v1

Goes, I. (2025). *When countries revise their data* [Working paper]. https://www.iasmingoes.com/papers/Goes-WDI.pdf

Groenendyk, M. (2023, January 4). *Tracking dataset citations and measuring data's research impact* [Conference presentation]. Hawaii International Conference on Education, Honolulu, HI, United States. https://hiceducation.org/wp-content/uploads/2022/12/23-final-program.pdf

Huang, J., & Chang, K. C.-C. (2023). *Citation: A key to building responsible and accountable large language models* (arXiv:2307.02185). arXiv. http://arxiv.org/abs/2307.02185

Istrate, A.-M., Li, D., Taraborelli, D., Torkar, M., Veytsman, B., & Williams, I. (2024). *A large dataset of software mentions in the biomedical literature* [Jupyter Notebook]. Chan Zuckerberg Initiative. https://github.com/chanzuckerberg/software-mentions (Original work published 2022)

Kratz, J. E., & Strasser, C. (2015). Making data count. *Scientific Data, 2*(1), Article 150039. https://doi.org/10.1038/sdata.2015.39

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology, 69*(11), 1346–1354. https://doi.org/10.1002/asi.24049

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the Data Citation Index. *Journal of the Association for Information Science and Technology, 67*(12), 2964–2975. https://doi.org/10.1002/asi.23529

Treasury Board of Canada Secretariat. (2020, March 31). *Open government implementation plan: Public Service Commission*. http://open.canada.ca/en/content/open-government-implementation-plan-public-service-commission

Vierkant, P. (2023, January 17). *Welcome Trust and the Chan Zuckerberg Initiative partner with DataCite to build the open global Data Citation Corpus*. DataCite. https://datacite.org/blog/data-citation-corpus-announcement-2023/

Zhang, Y., Wang, Y., Wang, K., Sheng, Q. Z., Yao, L., Mahmood, A., Zhang, W. E., & Zhao, R. (2023). *When large language models meet citation: A survey* (arXiv:2309.09727). arXiv. https://doi.org/10.48550/arXiv.2309.09727