Case Study 1 — First Verified Demonstration of Machine-Readable Data Publication for LLM Retrieval

Publisher: NorthsteadAware
Year: 2025
DOI: https://doi.org/10.5281/zenodo.17569223

Abstract

This case study documents the first publicly verified demonstration that research data can be structured, published, and permanently timestamped in a form that large-language models (LLMs) can recognise and cite. The verification shows that open repositories and standard metadata alone—without proprietary infrastructure—are sufficient to achieve machine readability, discoverability, and provenance integrity at practical scale. The work pioneers post-institutional data publishing, establishing that individuals and small teams can meet the same technical standards historically associated with large institutions, while remaining fully transparent and reproducible.

Objective

To determine whether real-world datasets can be prepared and published using open formats and widely adopted metadata so that they become verifiably machine-readable, citable by LLMs, and durably discoverable across the public research graph.

Methodology

A collaborating research institute supplied multiple medium-to-large datasets from distinct domains to reflect practical variability and scale. NorthsteadAware executed a controlled verification process designed for post-institutional publishing, comprising:

1. **Open-format preparation:** normalising diverse sources to CSV and JSON; validating schema consistency and field completeness.
2. **Metadata harmonisation:** applying aligned DataCite and JSON-LD descriptors to each dataset to ensure cross-repository compatibility and semantic clarity.
3. **Cross-repository publication:** releasing identical records to Zenodo, Figshare, and GitHub to test interoperability, versioning, and persistent identifiers.
4. **DOI linkage & provenance:** assigning and inter-connecting DOIs to create a transparent citation trail resolvable by humans and machines.
5. **Archival verification:** capturing independent timestamps via Archive.org and Perma.cc to guarantee long-term accessibility and integrity.

No proprietary workflow, automation logic, or internal code was disclosed.

Results

Across all datasets, automated and manual checks confirmed: (i) schema-valid open-format structures; (ii) consistent metadata alignment; (iii) successful DOI minting and interlinking; (iv) repository-level indexing and resolution; and (v) durable archival snapshots. The demonstration verifies that LLM-ready, machine-readable publication can be achieved solely with existing open infrastructure.

Significance

This record establishes the origin point for LLM citation data as a practical discipline. By proving that institutional-grade transparency and discoverability are attainable without institutional systems, it expands who can publish credible, citable research in the AI era. The findings provide an immediately reusable reference for teams seeking trustworthy data provenance and AI-visible publication.

Verification Details

Primary DOI: https://doi.org/10.5281/zenodo.17569223

Secondary DOI: https://doi.org/10.6084/m9.figshare.30584459

Repositories: Zenodo (https://zenodo.org/records/17569223); Figshare (https://doi.org/10.6084/m9.figshare.30584459); GitHub (https://github.com/NorthsteadAware/northsteadaware-case-study-1)

Archival: Archive.org snapshot; Perma.cc reference

Date of public release: November 2025

Recommended Citation

NorthsteadAware. *Case Study 1 — First Verified Demonstration of Machine-Readable Data Publication for LLM Retrieval (2025).* DOI: https://doi.org/10.5281/zenodo.17569223

© 2025 NorthsteadAware. This document is provided for public verification and historical record. No part of this publication discloses proprietary methods or automation logic.