DCPAPERS

DCMI 2026

Proceedings of the International Conference on Dublin Core and Metadata Applications 2026

ISSN: 1939-1366

Stop Guessing, Start Grounding: Using Metadata to Solve the Opacity Problem in AI Retrieval

Agnes Kleinhans[1], **Ning Xia**[1], **Andreas Noback**[1], and **Peter F. Pelz**[1]

[1]Technical University of Darmstadt

Abstract

freeda is an early-stage prototype AI information retrieval system that improves trust through transparent metadata. Its Explainability Layer reveals provenance paths, confidence scores, timestamps, and sources behind each answer. This approach shows how structured metadata can support more reliable and accountable AI use.

1. Introduction

As AI systems increasingly mediate access to information, the question of what metadata can do for human trust has become urgent. In an environment where misinformation spreads precisely because sources and certainty are obscured, making AI retrieval systems transparent is not a technical nicety but a public good. Existing research shows that transparency mechanism can reduce trust callibration error and improve appropriate reliance on AI outputs [1]. freeda (free data, free science, free society) is an early-stage prototype[1] that addresses this challenge directly. It is a natural language retrieval information system designed to surface fragmented research data from distributed institutional sources while making the reasoning behind each answer fully visible to users. Developed initially in the context of connecting research data across universities in the Rhein-Main region (Germany) freeda treats metadata transparency as a primary design requirement. Its central contribution is an *Explainability Layer* that grounds obtained answers in structured metadata (provenance paths, confidence scores, indication of source, and timestamps) giving users the tools to evaluate, verify, and act on retrieved information responsibly.

2. System Overview

freeda is built on LightRAG, a retrieval framework that constructs knowledge graphs over indexed source documents and traverses them at query time. Unlike vector matching, this graph-based approach enables reasoning over explicit entities and relations, making the system's inference chain traceable and its outputs anchored in structured metadata. [2] Users interact through a conversational web interface: Responses consist of an AI-generated answer (quick response), a set of source URLs indicating where the information was retrieved from, and an explainability layer. (see Figure 1).

[1]freeda can be explored as a prototype at: https://rmu-freeda.de

Poster

Available Aug 01, 2026

Correspondence to
Agnes Kleinhans
agnes.kleinhans@tu-darmstadt.de

3. Explainability Layer: Metadata in Service of Human Judgment

The Explainability Layer operationalises four categories of meaning-driven metadata, each addressing a distinct condition of trustworthy information use.

Knowledge graph provenance path. "How did the system reach this answer?" Each response displays the graph traversal that produced it—for example: Goethe University → has library → University Library → has biological collections → Louise von Panhuys [temporal coverage: 1763–1844]. Provenance paths let users trace and correct problematic sources or reasoning steps, supporting fairness and accountability in downstream use [3].

Confidence score. "Should I rely on this answer?" A numeric score reflects retrieval certainty based on the coverage and consistency of supporting graph nodes. Offering provenance-backed rationales alongside numeric uncertainty gives users both why and how much to trust an assertion, reducing both overtrust and undertrust [1]. Confidence scores also enable configurable acceptance thresholds, so users can determine when the system acts autonomously versus when human review is required [4].

Source timestamp. "How current is this?" Each response displays the date on which the underlying source data was indexed. Temporal metadata signals the freshness or staleness of evidence and supports time-aware filtering [5]. In research contexts, where temporal scope matters, making this visible by default treats currency as a baseline transparency requirement rather than an optional annotation.

Source links. "Where can I verify this?" Alongside each response, freeda surfaces the URLs from which information was retrieved. Direct source links make underlying documents immediately inspectable, supporting verification and contestability by users [6]. Combined with the provenance path, they close the loop between AI-generated answer and primary evidence.

4. Current Status and Next Steps

freeda is currently an early-stage proof of concept. The system has been demonstrated with digitized (botanical) collections held at Goethe University Frankfurt as well as person-related data from the Technical University Darmstadt. Planned next steps include expanding coverage across the Rhein-Main universities, developing a formal evaluation framework for retrieval accuracy and provenance completeness, and conducting structured user studies with researchers, students, and general public audiences.

5. Conclusion

freeda demonstrates that AI-powered retrieval and metadata transparency are complementary rather than competing goals. Explainability can be built into the retrieval architecture through structured metadata (provenance paths, source timestamps, indication of source, confidence scores) rather than added as an afterthought. As research data continues to fragment across platforms and institutions, systems that surface not

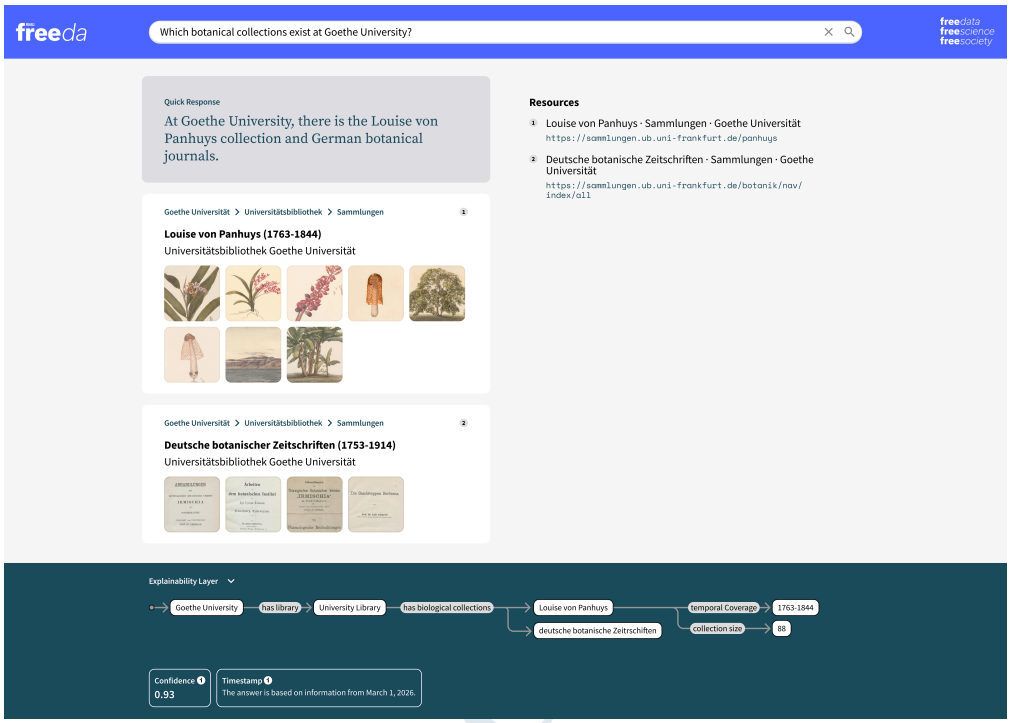

Figure 1. freeda *Interface featuring a metadata-driven Explainability Layer and Knowledge Graph Provenance path.*

just answers but the provenance of those answers will become increasingly important for trustworthy open science. freeda is an early step toward that vision, and this POSTER

References

[1] C. Newen, D. Bodemer, S. Glantz, E. Müller, M. Wischnewski, and L. Schnaubert, "Uncertainty Awareness and Trust in Explainable AI: On Trust Calibration using Local and Global Explanations," *CoRR*, vol. abs/2509.08989, 2025, doi: 10.48550/arXiv.2509.08989.

[2] Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, "LightRAG: Simple and Fast Retrieval-Augmented Generation," *CoRR*, vol. abs/2410.05779, 2024, doi: 10.48550/arXiv.2410.05779.

[3] B. Zhang, A. Meroño-Peñuela, and E. Simperl, "Towards Explainable Automatic Knowledge Graph Construction with Human-in-the-Loop," *Frontiers in Artificial Intelligence and Applications*, vol. 372, pp. 274–289, 2023, doi: 10.3233/FAIA230091.

[4] S. Hani *et al.*, "When to Trust the Machine: A Simulation Framework for Human–AI Collaboration," *Advances in Human Factors and Ergonomics*, vol. 200, 2026, doi: 10.54941/ahfe1007095.

[5] T. Huang and E. Fan, "Structured Reasoning for Fairness: A Multi-Agent Approach to Bias Detection in Textual Data," *CoRR*, vol. abs/2503.00355, 2025, doi: 10.48550/arXiv.2503.00355.

[6] R. Souza *et al.*, *LLM Agents for Interactive Workflow Provenance: Reference Architecture and Evaluation Methodology*. in Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2025, pp. 2257–2268. doi: 10.1145/3731599.3767582.