

Non-Archival Abstract

Mind the Gap: Bridging Query Formulation and Provider Vocabularies in Research Data Registries

Timm Lehmberg

Akademie der Wissenschaften Hamburg

This poster covers two approaches to LLM-enhanced research data retrieval in the context of the federated research data infrastructure Text+. As part of the National Research Data Infrastructure (NFDI), the German Text+ consortium federates over thirty data centers, providing language resources from three data domains: collections, (scholarly) editions, and lexical resources. In order to make these resources discoverable, the Text+ Registry, in its function as central catalogue, exposes a unified metadata-based search across the entire spectrum of these holdings.

However, both in the process of curating metadata as well as querying the aggregated information, problems arise that can hardly be adequately addressed by traditional information retrieval approaches.

For example, it emerged that data providers were puzzled as to why their resources surface incorrectly or with poor ranking. On the users' side in turn, researchers from various disciplines often face a persistent query-formulation gap, where vague queries, non-standard terminology, and idiosyncratic provider vocabularies routinely produce empty or misleading result sets.

Therefore, the two prototypes presented in the following (a GraphRAG-based discovery layer and an LLM-augmented hybrid retrieval interface) try to enrich the retrieval process with semantic and pragmatic context, addressing the two sides of the problem from different directions.

RAGestry: GraphRAG-Mediated Discovery

The Text+ RAGestry addresses the query-formulation gap from the user side. Following the GraphRAG paradigm, it extends retrieval-augmented generation through the integration of a knowledge graph layer, allowing knowledge to be extracted from larger, cross-textual contexts. Metadata and full-text descriptions from the Text+ Registry are ingested in two stages: structured fields containing largely unambiguous information are modelled and mapped into a knowledge graph alongside a vector store; free-text descriptions are then enriched through multilingual sentence transformers and a generative LLM, with no pre-specified entity constraints, so that semantic relationships are discovered automatically through the content itself. The first prototype handles keyword and natural-language queries alike, including multilingual, terminologically idiosyncratic, and misspelled formulations.

Registry Searchlab: Hybrid Retrieval with User Feedback

The Searchlab addresses the related query-formulation gap from the opposite (provider-) side, exposing the mechanics of retrieval to data providers, who frequently lack visibility into why their resources do or do not surface in aggregated views. Each query is processed in parallel by a multilingual sentence-transformer model and a BM25 full-text engine. Cosine and lexical scores are fused with per-resource-type weighting that reflects metadata heterogeneity across the three Text+ domains. On demand, a locally hosted generative LLM produces a human-readable explanation for any result, separating semantic from lexical matching signals and flagging low-information entries. In addition, a dedicated feedback index captures thumbs-up / thumbs-down ratings and explicitly named missed documents, dynamically boosting, demoting, or injecting items in subsequent queries. Beyond this, data curators control fusion parameters through interactive sliders, tuning retrieval behaviour in real time without code changes.

References

- Barth, F., Draxler, C., Ecker, J., Fischer, S., Genêt, P., Hemmer, A., Lehmberg, T., Trippel, T., Witt, A., Zimmermann, A. & Zinn, C. (2026). Text+: A National Hub Including Legacy Language Data. In *Proceedings of LREC 2026*, pp. 8264–8275. <https://doi.org/10.63317/4vx5d59r6m29>
- Grädl, T., Fruth, L. & Henrich, A. (2025). The Text+ Registry: Federating Research Data Catalogues for the Digital Humanities. *TPDL 2025*, LNCS 16097.

- Guo, Z., Xia, L., Yu, Y., Ao, T. & Huang, C. (2025). LightRAG: Simple and Fast Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10746–10761.
- Larson, J. & Truitt, S. (2024). *GraphRAG: Unlocking LLM discovery on narrative private data*. Microsoft Research Blog.
- Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS 2020*.