Practical Prompt Engineering for AI-Supported Metadata Processing in CRIS

Danica Zendulková, Slovak Centre of Scientific and Technical Information, Lamačská 8/a, 840 05 Bratislava, Slovakia, danica.zendulkova@cvtisr.sk, ORCID: 0000-0003-2487-0177

Otmane Azeroual
University of Hagen, Universitätsstraße 47, 58097 Hagen, Germany
otmane.azeroual@fernuni-hagen.de, ORCID: 0000-0002-5225-389X

Type of contribution: research paper

CRIS 2026 conference call for papers
https://eurocris.org/cris2026-conference-azores-call-proposals

Extended abstract

The growing digitalization of research management has transformed the availability, quality, and interoperability of research information into a central strategic factor within contemporary science systems. Research institutions are increasingly confronted with extensive volumes of publications, projects, datasets, and related outputs that must be systematically captured, classified, curated, and disseminated across national and international research information infrastructures (Fabre et al., 2021). Current research information systems (CRIS) have therefore become indispensable institutional components (Sivertsen, 2019). At the same time, evolving policy expectations around Open Science, research assessment reform, the FAIR data principles, and evidence-based institutional governance have amplified the requirements for consistent, transparent, and interoperable metadata (Wilkinson et al., 2016; Wilsdon et al., 2015; Azeroual et al., 2022). Furthermore, the integration of AI-driven analytics into CRIS can enable real-time monitoring of research activities and dynamic identification of emerging scientific trends, providing institutions with predictive insights for strategic planning (Azeroual, 2025a).
Despite these increasing demands, metadata workflows in many organizations remain highly manual and heterogeneous, relying heavily on individual interpretation and localized convention. Research in information science has long identified metadata quality as a persistent bottleneck in digital knowledge infrastructures (Bruce &

Hillmann, 2004; Park, 2009; Borgman et al., 2015). This challenge is exacerbated in the CRIS context, where metadata no longer serve merely descriptive purposes but underpin strategic decision-making, reporting to funders, international benchmarking, and institutional visibility. Manual assignment of scientific classifications, keywords, subject domains, or project categories is time-consuming, prone to inconsistency, and difficult to scale. Empirical studies repeatedly show that a lack of standardized procedures and varying interpretive practices contribute significantly to data heterogeneity and reduced metadata reliability (Leiva-Mederos et al., 2017; Azeroual & Schöpfel, 2019; Zeng & Qin, 2020; Zendulková et al., 2025). Simultaneously, most institutions lack dedicated AI teams or sufficient training data to develop custom machine-learning models capable of automating such tasks. Moreover, cross-institutional data harmonization and alignment with global ontologies remain largely unexplored, representing an opportunity to establish interoperable metadata networks at regional and international levels.

Recent advances in large language models (LLMs) offer a new, low-threshold approach: the systematic use of prompt engineering to generate high-quality metadata through text-based instruction alone (Zhang et al., 2025). LLMs have demonstrated robust performance in tasks requiring text understanding, semantic classification, and domain mapping (Gatto et al., 2025; Gilardi, Alizadeh, & Kubli, 2023; Azeroual, 2025b). The approach presented in this study operationalizes these emerging capabilities specifically for CRIS environments by embedding institutional classification rules, controlled vocabularies, and documentation standards directly into structured prompts. This enables LLMs to generate consistent, explainable, and reproducible metadata suggestions that are subsequently validated by expert practitioners. Additionally, the workflow incorporates iterative feedback loops to continuously refine AI performance, enabling adaptive learning from human corrections and emerging research patterns. The architecture of this workflow, integrating machine assistance with human oversight, is illustrated in Figure 1.

Figure 1. Solution Pipeline.

This hybrid human-in-the-loop model aligns with established principles of responsible AI use in sensitive data domains (Shneiderman, 2020). It reduces manual workload, increases classification consistency by enforcing uniform rule application, enhances transparency through explicit reasoning steps, and can be deployed without modifying existing CRIS infrastructures. As a result, institutions without specialized technical staff or extensive training datasets can still access high-quality, AI-supported metadata processing. Furthermore, this approach opens avenues for collaborative networks of CRIS systems that share AI-enhanced metadata, creating a federated, interoperable research knowledge graph at the national or international level (Fabre & Azeroual, 2024). Overall, prompt engineering emerges as a scalable, transparent, and institutionally practical method for overcoming long-standing challenges in metadata workflows, improving the governance, interoperability, and strategic value of research information across the scholarly ecosystem.

Building on this conceptual foundation, the paper focuses on concrete application scenarios and hands-on examples demonstrating how artificial intelligence and prompt engineering can be used to support the processing of research project metadata in CRIS environments.

The study explicitly addresses the operational challenges faced by CRIS administrators and data curators in real-world institutional settings.

We start from a situation in which the CRIS system administrator is faced with the task of inserting a large amount of data about research projects from an external environment into the system. This data does not come from an external information system and therefore it is not possible to use the import API interface of the CRIS system. The data is usually recorded in MS Excel and does not correspond exactly to the database structure of the CRIS system. Some fields are not filled in correctly or at all, therefore the data does not meet the validation criteria. Checking and supplementing data is therefore a necessary step that must be taken before uploading them to the system.

This scenario represents a common and often underestimated bottleneck in CRIS data integration workflows, particularly in smaller institutions or decentralized research environments.

The paper demonstrates how LLM-based prompt engineering can support curators in validating, enriching, harmonizing, and restructuring such semi-structured datasets prior to system ingestion.

The aim of the paper is to prove the hypothesis that currently commonly available artificial intelligence tools in free versions can speed up and facilitate the work of data curators in preparing and checking data for the CRIS system.

More specifically, the study investigates whether freely accessible LLM-based tools can significantly improve efficiency, consistency, and data quality in CRIS metadata preparation without requiring programming expertise or system-level integration.

Keywords

Research Information Systems (CRIS), Metadata Quality, Prompt Engineering, Large Language Models (LLMs), Automated Metadata Classification, Human-in-the-Loop.

References

Azeroual, O., & Schöpfel, J. (2019). Quality issues of CRIS data: An exploratory investigation with universities from twelve countries. *Publications*, *7*(1), 14.

Azeroual, O., Schöpfel, J., Pölönen, J., & Nikiforova, A. (2022). Putting FAIR principles in the context of research information: FAIRness for CRIS and CRIS for FAIRness. In *14th International Conference on Knowledge Management and Information Systems (KMIS2022)* (pp. 63-71). SCITEPRESS-Science and Technology Publications.

Azeroual, O. (2025a). AI4Knowledge: Shaping the Future of Research Data Systems. In *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)* (pp. 1-7). IEEE.

Azeroual, O. (2025b). Streamlining Research Information Management with Large Language Models: Automation, Analysis, and User-Centric Solutions. In *10th International Conference on Big Data Analytics (ICBDA)* (pp. 281-287).

Bruce, T. R., & Hillmann, D. (2004). The continuum of metadata quality: Defining, expressing, exploiting. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 1–8, ALA editions.

Fabre, R., Egret, D., Schöpfel, J., & Azeroual, O. (2021). Evaluating the scientific impact of research infrastructures: The role of current research information systems. *Quantitative Science Studies*, *2*(1), 42-64.

Fabre, R., & Azeroual, O. (2024). Knowledge graphs–the future of integration in CRIS systems for uses of assistance to scientific reasoning. *Procedia Computer Science*, *249*, 264-279.

Gatto, J., Sharif, O., Seegmiller, P., Bohlman, P., & Preum, S. (2023, December). Text encoders lack knowledge: Leveraging generative llms for domain-specific semantic textual similarity. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)* (pp. 277-288).

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120.

Leiva-Mederos, A., Senso, J. A., Hidalgo-Delgado, Y., & Hipola, P. (2017). Working framework of semantic interoperability for CRIS with heterogeneous data sources. *Journal of Documentation*, *73*(3), 481-499.

Park, J. R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & classification quarterly*, *47*(3-4), 213-228.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504.

Sivertsen, G. (2019). Developing Current Research Information Systems (CRIS) as data sources for studies of research. In *Springer handbook of science and technology indicators* (pp. 667-683). Cham: Springer International Publishing.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., ... & Johnson, B. (2015). The metric tide. *Report of the independent review of the role of metrics in research assessment and management*. Higher Education Funding Council for England.

Zhang, G., Wang, X., Sun, S., Xiao, P., Sun, K., & Xiong, Y. (2025). TraceRAG: A LLM-Based Framework for Explainable Android Malware Detection and Behavior Analysis. *arXiv preprint arXiv:2509.08865*.

Zendulková, D., Azeroual, O., & Krivjanská, A. (2025). From Data Silos to FAIR Research: Interoperability and Integration in European CRIS. In *Digitalization and Management Innovation IV* (pp. 228-240). IOS Press.

Zeng, M. L., & Qin, J. (2020). *Metadata*. American Library Association.