Metadata Extraction with LLMs for DCAT-AP+ based Research Data Management in Process Engineering and Catalysis

Marc Völkenrath, TU Dortmund University, Dortmund/Germany; Simon Clemens, TU Dortmund, Dortmund/Germany; Hendrik Borgelt, TU Dortmund, Dortmund/Germany; Dr.-Ing. Alexander S. Sommer-Behr, TU Dortmund University, Dortmund/Germany; Prof. Dr.-Ing. Norbert Kockmann, TU Dortmund University, Dortmund/Germany;

Efficient and FAIR (Findable, Accessible, Interoperable, Reusable) research data management is essential for sustainable data reuse and reproducibility in catalysis research and chemical engineering. Heterogeneous data sources, inconsistent documentation practices, and insufficiently standardized metadata continue to complicate semantic interoperability and long-term accessibility of experimental data. Within the “Nationale Forschungsdateninfrastruktur” (NFDI) this work presents a workflow that automates the extraction, validation, and semantic enrichment of metadata assisted by Large Language Models (LLMs) from scientific datasets in various file formats.

The workflow applies a customized Ollama-LLM [1] combined with the DCAT-AP+ [2] metadata schemas and the Voc4Cat [3] domain vocabulary. Relevant domain concepts are identified through lexical and semantic matching and assigned to schema-compliant metadata structures. Missing concepts are detected using definition-based reasoning with existing vocabularies and ontologies [4]. If no suitable matches are found, the LLM proposes new, standard-compliant candidate concepts, which are subsequently reviewed and validated by domain experts to ensure semantic correctness and consistency.

The user validated metadata are exported in a standardized, machine-readable representation compatible with existing research data infrastructures. The feedback of domain experts is used to extend the metadata schemas and the controlled vocabularies and ontologies. The resulting enriched metadata and semantic resources enable the construction of interoperable knowledge graphs that can be validated using SHACL and queried via SPARQL to support enhanced literature search, improved research planning and knowledge discovery.

Acknowledgements:

The NFDI4Cat Project (NFDI 2/2,441926934) is acknowledged for financial support of this work.

References:

[1] https://ollama.com/ (Accessed on September 30, 2025).

[2] P. Strömert, H. Borgelt, M. Doerr, D. Linke, *nfdi-de/dcat-ap-plus: Release 0.1.0rc3 (same as rc2)*, Zenodo **2025**.

[3] D. Linke, N. Moustakas, M. Doerr, J. Schumann, M. Götte, H. Borgelt, M. Nentwich, C. Terboven, F. Flecken, *nfdi4cat/voc4cat: Release 2025-10-14*, Zenodo **2025**.

[4] A. S. Behr, H. Borgelt, N. Kockmann, *Journal of cheminformatics* **2024**, *16 (1)*, 16. DOI: https://doi.org/10.1186/s13321-024-00807-2