Facilitating Comprehensive Metadata Capture and Validation in Data Repositories (nmrXiv) through Terminologies and Terminology Service Suite Widgets

Authors:
Venkata Nainala[1]*, Noura Rayya[2], Christoph Steinbeck[2], Oliver Koepler[3]
[1] chandu.nainala@uni-jena.de, Friedrich Schiller University, Jena, Germany
[2] Friedrich Schiller University, Jena, Germany
[3] TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany

Abstract:

Analytical chemistry, one of the oldest scientific disciplines, integrates methods from physical, inorganic, and organic chemistry. The standard workflow begins with formulating a research question and devising experiments, methodologies, and surveys to assess the hypotheses. Experiments are conducted, and samples are treated using established or novel methods, with the resulting data being meticulously recorded. However, the metadata captured during these experiments often needs more details for reproduction and is frequently not machine-readable. A survey conducted by Herres-Pawlis et al. revealed that only 42% of participants described their collected data with metadata [1]. This partial or nonexistent annotation leads to misinterpretation and the loss of time and resources, ultimately hindering future research progress.

Data repositories face the challenge of enabling user-friendly yet high-quality metadata capture. Ensuring that metadata is both comprehensive and easy for researchers to input requires robust terminologies and seamless integration of annotation tools. In this era of large language models (LLMs), the quality of data is paramount, as LLMs rely heavily on high-quality, well-annotated data to generate accurate and reliable insights. Currently, analytical data repositories encounter significant limitations in data annotations, primarily due to insufficient or incorrect terms available in existing terminologies. This results in missing or inaccurate annotations, compromising the quality and reliability of the metadata. For instance, the ontology nmrCV [2], essential for NMR metadata, has several issues that require attention. Domain-independent problems include structural inconsistencies, incorrect annotations, incomplete definitions, and outdated documentation. Domain-specific problems pertain to the NMR field and include inaccuracies in definitions, misaligned hierarchies, and the inclusion of non-NMR entities. To mitigate these issues, collaborative efforts with NFDI4Chem TA6 have been undertaken to enhance the terminologies. This work focuses on improving properties, annotations, and documentation, specifically concerning NMR solvents, calibration compounds, and instrument manufacturers. The goal is to ensure that the ontology accurately represents NMR metadata and aligns with other specialised ontologies such as CHEBI.

On the other end, the lack of effective integration of rich, terminology-driven form widgets within data repositories exacerbates the issue. The nmrXiv data repository, developed as part of the NFDI4Chem initiative, employs the ontology data of the NFDI4Chem terminology service [3, 4] during the data provisioning phase via widgets provided by the Terminology Service Suite [5] from the Terminology Services 4 NFDI (TS4NFDI) base service. This integration allows

submitters to annotate their data with metadata in a semi-automatic manner, significantly enhancing both metadata quality and user experience. Additionally, this enables nmrXiv to actively learn from these new annotations by engaging with data submitters. This iterative process will refine the knowledge graph, with the ultimate goal of achieving fully automated metadata annotation.

Keywords: Metadata Annotation, Analytical Chemistry, NMR Ontology, Terminology Services, Data Repositories

[1] Herres-Pawlis, S., Koepler, O., & Steinbeck, C. (2019). NFDI4Chem: Shaping a Digital and Cultural Change in Chemistry. Angewandte Chemie International Edition, 58(32), 10766-10768. https://doi.org/10.1002/anie.201907260
[2] FAIRsharing.org: nmrCV; Nuclear Magnetic Resonance Controlled Vocabulary, DOI: 10.25504/FAIRsharing.xm7tkj
[3] Strömert, P., Limbachia, V., Oladazimi, P., Hunold, J., Koepler, O., Towards a versatile Terminology Service for empowering FAIR research data: Enabling ontology discovery, design, curation, and utilization across scientific communities. Studies on the Semantic Web, Vol. 56 Knowledge Graphs: Semantics, Machine Learning, and Languages. IOS Press; 2023. doi:10.3233/ssw230005
[4] https://terminology.nfdi4chem.de
[5] https://github.com/ts4nfdi/terminology-service-suite