A Semantic Laboratory Assistant for Metadata Acquisition in Electronic Lab Notebooks

Abstract

Laboratory data reuse and reproducibility depend on rapid, accurate, and complete capture of experimental (meta)data. In practice, metadata creation in electronic lab notebooks (ELNs) remains a bottleneck because form-based entry interrupts workflows, free-text input is time-consuming and error-prone, and heterogeneous terminology complicates harmonisation across projects and infrastructures. These limitations reduce metadata quality and quantity and impede Findable, Accessible, Interoperable, and Reusable (FAIR) dissemination and integration. LabFriend is an open, ELN-agnostic laboratory assistant under development to mitigate these issues through semantically structured, context-aware metadata acquisition. Intended functionality includes real-time suggestions, validation of field values against controlled semantics, and optional speech-based capture. Methods combine association-rule mining from historical form instances with ontology- and knowledge-graph-based semantic relatedness, aiming to improve completeness and terminology consistency while keeping interaction lightweight.

A central prerequisite is robust data preparation that converts heterogeneous ELN exports into validation-ready material for semantic methods and evaluation. This contribution focuses on a preparation workflow for transforming records collected in the Chemotion ELN into knowledge-graph-ready representations. The workflow addresses common obstacles in exported records, including a mixture of structured key-value fields and unstructured free-text, missing or implicit units, inconsistent naming, ambiguous identifiers. Preparation steps include structure extraction from Chemotion objects, normalisation of datatypes and units, entity resolution across samples, processes, and instruments, and semantic anchoring to domain vocabularies while preserving provenance of each transformation decision. The resulting material is annotated manually against a closed, schema-driven target model and can be mapped to Resource Description Framework (RDF) statements for knowledge-graph construction and downstream reuse.