Contribution ID: **9** Type: **not specified**

AI-assisted data annotation for biomedical research consortia

Tuesday, September 30, 2025 2:15 PM (1h 30m)

Accurate and complete metadata is key for making research data findable and thus potentially reusable. However, since data annotation is a tedious and time-consuming process, compliance is often low, leading to missing or incomplete data set descriptions. A possible solution to this problem is the automatic extraction of metadata from data files and associated documents such as research articles and laboratory notebooks. The recent advancement in development of Large Language Models (LLMs) promises to greatly facilitate automatic data annotation, since LLMs excel in extracting specific information from text documents and inferring related concepts from the information provided in the text.

Here we present an approach for LLM-based automatic research data annotation developed by the Research Data Management (RDM) group at the Institute of Medical Biometry and Statistics (IMBI) of the University Medical Center Freiburg. The IMBI RDM group currently supports seven biomedical research consortia by offering training, consulting and software solutions for RDM. Consortium members have access to the open-source electronic laboratory notebook eLabFTW as well as the in-house developed research data platform fredato (Freiburg research data tool). fredato is a platform for collaboration and data sharing built on the open-source components Nextcloud (file storage, sharing and collaborative editing), GitLab (version control and metadata storage) and OpenSearch (search engine). To allow scientists to easily annotate their data with terms most relevant to their use case, the RDM group develops a metadata schema for each individual consortium in close collaboration with the scientists. This metadata schema is implemented as a JSON schema from which a web form for data annotation is built. Metadata are stored as JSON files and can be exported to public data repositories. For interoperability, elements of the metadata schemas are linked to biomedical ontologies and the metadata schemas are published as RDF (Resource Description Framework) knowledge graphs.

While bibliographic metadata of biomedical research articles can be imported from the PubMed database, data sets currently need to be annotated manually by the authors. To minimize the time and effort required for data annotation, we have developed a workflow of metadata prediction from publication PDFs using LLMs. The four-step workflow identifies biomedical entities (e.g. organisms, cell lines, genes, diseases) in the article full text using ChatGPT in combination with the consortium's metadata schema and the PubTator3 database of biomedical entities. We tested this workflow with 23 publications of the Collaborative Research Center 1479 "OncoEscape" and validated the suggested terms in structured interviews with publication authors. Overall, the approach worked very well with a precision of 98% (95% CI: 94%-100%), i.e., the vast majority of predicted biomedical entities were considered correct in the face-to-face interviews. Including or excluding the article's supplementary material did not result in a significant difference in precision of suggested terms.

We are currently working on implementing the LLM-assisted metadata prediction in fredato and testing the performance of several alternative LLMs. Thus, in the near future researchers will only need to review the pre-filled metadata web form and delete incorrect entries. In addition, we are testing the use of LLM-assisted metadata prediction in the schema development process to identify potentially overlooked terms.

Finally, we plan to extend the LLM-assisted metadata prediction to other sources beside article PDFs, including electronic laboratory notebooks such as eLabFTW and files stored on the OMERO platform for microscopy data. This will allow us to capture metadata directly in the process of data creation, thereby ensuring that research data are immediately findable by interested researchers.

Abstract

Poster

Author: Dr BENADI, Gita (Institute of Medical Biometry and Statistics, Medical Faculty and University Medical Center, University of Freiburg)

Co-authors: Dr GIULIANI, Claudia (Institute of Medical Biometry and Statistics, Medical Faculty and University Medical Center, University of Freiburg); Dr ENGEL, Felix (Institute of Medical Biometry and Statistics, Medical Fac-

ulty and University Medical Center, University of Freiburg); Prof. BINDER, Harald (Institute of Medical Biometry and Statistics, Medical Faculty and University Medical Center, University of Freiburg); Dr KAIER, Klaus (Institute of Medical Biometry and Statistics, Medical Faculty and University Medical Center, University of Freiburg)

Session Classification: Poster Session