Helping Ontology Extension with Natural Language Processing for Catalysis

Ontologies store semantic knowledge in a machine-readable way and represent domain knowledge in controlled vocabulary. Scientific results often are published in text form, thus discouraging research data FAIRness. Using natural language processing (NLP), concept names and relations can be extracted from text datasets.

A workflow to process scientific textual text corpora is introduced regarding catalysis research ~~is introduced~~. NLP techniques are used to vectorize the textual data. This allows for hierarchical clustering of concepts, also yielding concept names. In addition, ontologies containing the resulting concept names are searched from a database. Once found, corresponding existing definitions of those concepts are also important output ~~allowing for~~enabling domain experts to validate correctly found ontology classes. Subsequently performed hierarchical clustering of the concept names based on the text corpora prepares the found data for ontology matching, assisting in ontology extension.

Previously undefined concepts and unstructured relations can thus be more easily introduced into existing ontologies based on their descriptive scientific texts. A structured extension of ontologies supported by NLP methods is thus made possible to facilitate FAIR data management workflow. The contribution shows successful applications and highlights existing hurdles, too.