1st Conference on Research Data Infrastructure

Linking RDM

https://doi.org/10.52825/CoRDI.v1i.310

© Authors. This work is licensed under a Creative Commons Attribution 4.0 International License

Published: 07 Sept. 2023

Harmonized research information for classifying and linking research data

Fadwa Alshawaf[1], Rolf Guescini[1], Florian Kotschka[1], Maik Bierwirth[1], and Malte Dreyer[1]

[1] Humboldt University of Berlin, Germany

Motivation

A well-described, harmonized in standard formats and linked metadata accelerates and refines the search process. Hence, researchers can ensure that their data is properly organized, easily discoverable, and accessible to others. Machine-readable metadata is especially essential for automatic discovery of research datasets and outputs. It facilitates the search and retrieval of data with increased accuracy, which saves the user time in finding crucial information within their field of research. This requires rich and categorized metadata that allows computers to automatically retrieve and sort relevant search results. Well documented metadata which is harmonized, standardized and serialized in linked data formats increases not only the accuracy of the search results within a field of research, but can also provide relevant and helpful suggestions within disciplinary and interdisciplinary fields of research.

Research information platform with VIVO

Within the framework of the Berlin University Alliance, we are creating a platform for structured capturing and transparent presentation of research information using the open-source software VIVO. The platform comprises a database and a frontend. The database of research information is linked to expert portfolios to showcase research expertise and activities, improve the discoverability of expertise, connect researchers to their work across disciplines and boundaries, as well as facilitate new research collaborations.

Since the project serves different institutions and collects data across disciplines, it is of significantly important to standardize the format of the collected research information. To achieve this, we represent and serialize the research information according to standardized linked data formats. We also classify the data under vocabularies such as those suggested with the project of interdisciplinary Research Core Dataset (Kerndatensatzforschung, KDSF) and standard vocabularies such EUDAT-B2Find and DESTATIS-subjects. This creates a comprehensive information model for heterogeneous scientific systems and supports the interchange of research information within the Alliance.

Metadata extraction and classification using NLP

As part of the project, a faster, more cost-effective approach is developed to automatically extract and analyze research information from websites or files of research outputs. Machine learning Natural Language Processing (NLP) technologies make it possible to analyze huge amounts of texts on web pages or in documents and automatically extract research information. Texts are automatically scanned and analyzed to extract entities such as names and

organizations and specific information such as research subject and area or predefined keywords and vocabularies.

Text classification enables automatic categorization of the extracted information under predefined tags or groupings. It also allows texts to be categorized by their context without predefined categories being explicitly present in the text. Still, the use of keywords and controlled vocabularies as part of the metadata is essential to classify the research data under subject-relevant categories, improve, and fasten the search results. Therefore, it is recommended to set of predefined categories and vocabularies for the data classification. This includes research disciplines, subject areas, and interdisciplinary research fields. These categories are often arranged in information architectures called taxonomies, which can be further converted into machine-understandable ontologies.

The aim of creating an ontology is the development of knowledge graphs, which build the foundation for intelligent systems that can understand, interpret, categorize, and link research artefacts based on natural language. In order to achieve an automatic classification of the research data, neural networks are trained and created through machine learning approaches and applied to the text shared on websites or in documents to classify and generate the metadata. Moreover, a network of ontologies is developed to link different scientific entities such as organizations, publications, datasets, and persons to improve the discoverability of research datasets and outputs.

This approach facilitates the finding and browsing of relevant research information and quick access to research that is done at an institute under a specific discipline or area of research. It could, for example, provide necessary information for strategic planning of future development and research collaboration. It also provides relevant suggested results in the context of the searched keywords.

Outlook

We will create a platform that emphasizes research projects and links them to other research entities such as organizations, publications, datasets, funding, and events. This directs the attention to the endeavors of promoting interdisciplinary, across-institutions, and international collaborations.

Future planning within the approach of metadata classification include extended classification methods as well as ontologies. This can support broader and further classification schemes as well as multilingual search possibilities to enable a greater spectrum of findability, accessibility, interoperability and reusability of the research data.