Aim of the team project

Study programme: Computer Science - Software and Data Engineering

Project type: Research Project (NPRG070)

Student: Martin Gora

Supervisor: Martin Nečaský

Consultant: Jakub Klímek, Petr Škoda, Štěpán Stenchlák

Project name: Integration of the Wikidata ontology into the Dataspecer tool

Design of the technical solution, time schedule, milestones and additional information:

Motivation:

A research team M. Nečaský, J. Klímek, P. Škoda a Š. Stenchlák aims to connect worlds of data structures and formats used for representation and exchange of data on the Web. Such formats can be CSV, XML, JSON or various RDF serializations. Recently, the team created a Dataspecer tool that provides a strict separation of the conceptual model from the target data structures. The Dataspecer tool currently enables users to create data structures with the support of "slovník.gov.cz" ontology and various RDF vocabularies. This project aims to extend the Dataspecer tool with the Wikidata ontology. Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others. The Wikidata ontology is an intrinsic part of the data provided by Wikidata. However, the accessibility and usage of the Wikidata ontology is hindered by the general quality and large volume of the provided data. The main part of this project is to design and implement integration of the Wikidata ontology into the Dataspecer tool and enable its usage to users during the data structure creation process.

Project description:

The student will join the aforementioned research team to design and implement integration of the Wikidata ontology into the Dataspecer tool that is being developed by the research team.

Firstly, the student will perform an analysis of the Wikidata data models and the access methods to Wikidata. Since Wikidata is frequently in the aim of research from the Wikidata community, research of existing methods and tools for processing and manipulating the Wikidata ontology will be conducted.

The principal part of the research project includes a development process divided into two steps: extraction and integration.

- Extraction
    - The student will choose methods of the ontology extraction based on the previous analysis of the Wikidata access methods. The two of the main access methods are SPARQL Endpoint and RDF/JSON dumps. The timeout of queries while using SPARQL Endpoint must be regarded. To tackle the timeouts, a LinkedPipes: ETL tool can be used.
    - It must be defined what parts of Wikidata will be used to create the resulting ontology.
    - The extraction must be repeatable, since Wikidata is periodically updated by the Wikidata community.
    - It must be decided if the creation of a backend service is needed to periodically extract and access the Wikidata ontology.
    - The Wikidata ontology contains errors (such as: cycles in hierarchy, unused classes, invalid links, …). The errors must be considered while extracting the ontology.
- Integration
    - The student will choose how the extracted Wikidata ontology is integrated into the Dataspecer tool.
    - The integration process will include extensions of the Dataspecer tool:
        - The user interface must support usage of the integrated Wikidata ontology during a data structure creation process.
        - If the backend service is created, the Dataspecer tool must allow connection to the service.
    - As of 8.8.2023, there are 105,781,358 existing items in the Wikidata. Since any item can contain a unique class definition (2,899,745 class definitions as a minimum), the usage of the Wikidata ontology in the Dataspecer tool must be performant to provide ease of use to the users.

The work will be done in an iterative fashion. Firstly, a straightforward solution will be implemented. Based on the outputs of the first solution, a design and an implementation of one or more elaborate solutions will be done. The most fitting solution will be enhanced into the final solution. The output of each iteration will be consulted with the research team members.

Project output:

The Dataspecer tool will support the Wikidata ontology as part of a data structure creation process.

Platforms, technologies:

- Typescript, React.js, Python, Java, C++
- SPARQL, RDF, JSON
- LinkedPipes: ETL

Milestones and time schedule:

Name of a milestone	**Time schedule**
An analysis of Wikidata data models and research of existing tools and methods for processing and manipulating the Wikidata ontology.	M01
1. iteration: A design and an implementation of one or more straightforward integration solutions of the Wikidata ontology into the Dataspecer tool.	M02
2. iteration: Stemming from the output of the previous iteration, an architecture design of an elaborate integration solution and a subsequent implementation.	M03-M04
3. iteration: Performance and data quality optimisations of the implemented solution to provide pleasant user experience while working with the Dataspecer tool.	M05-M07
An output of each iteration encompasses consultations about reached aspects and attributes of the implemented solutions that serve as improvement proposals for the sequential iteration.	M02-M07
A finished documentation.	M08
The submission of the project.	M09