Representing Jupyter Notebooks with Knowledge Graphs to Address Data Lineage Problems

Auteur : Birtles, Alixia
Promoteur(s) : Debruyne, Christophe
Faculté : Faculté des Sciences appliquées
Diplôme : Master : ingénieur civil en science des données, à finalité spécialisée
Année académique : 2023-2024
URI/URL : http://hdl.handle.net/2268.2/20479

University of Liège
School of Engineering
Academic year 2023-2024

Author: Birtles Alixia
Supervisor: Debruyne Christophe

Representing Jupyter Notebooks with Knowledge Graphs to Address Data Lineage Problems

Master Thesis carried out to obtain the degree of Master in Data Science and Engineering

In data science, data lineage is a crucial aspect that is often insufficiently considered. To address challenges related to data lineage, the approach presented in this thesis leverages knowledge graphs and data provenance.

The PROV-O ontology and the FOAF vocabulary are harnessed to design a structure, along with defined terms. This ontology aims to represent the information extracted from Jupyter notebooks, tools often used in data science. Additionally, public APIs are leveraged to enrich the graph.

Initially, the RML language was used to map the data, but it was too limiting and led to the consideration of the RDFLib library in Python. RMLMapper and Morph-KGC have been considered, but the former does not have the required extension to access the desired data in the source code, while the latter has iterator challenges and does not support theta-joins.

The correctness of the approach was validated with visualization in GraphDb and SPARQL queries. A complex query related to the extraction of licenses demonstrated the feasibility of the approach and the ability to answer questions about data lineage. Moreover, experimentation with queries on a real-world dataset, the KGTorrent dataset, showed the effectiveness of the approach. Performance measurements on the construction of the graph and on SPARQL queries in real-world conditions led to promising results.