
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>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

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.