Opening up Participation in the Semantic Web

A New Data Upload Pipeline

Lozana Rossenova
Open Science Lab
German National Library of Science and Technology
Hannover, Germany
lozana.rossenova@tib.eu

Lucia Sohmen
Open Science Lab
German National Library of Science and Technology
Hannover, Germany
lucia.sohmen@tib.eu

EXTENDED ABSTRACT

In this short paper we will present a pipeline workflow developed in the context of NFDI4Culture, a German consortium of research and cultural institutions working towards a shared infrastructure for research data that meets the needs of 21st century data creators, maintainers and end users. The pipeline workflow has an explicit focus on making data curation accessible, promoting FOSS and FAIR principles, and collaborating with Wikimedia Germany, an official partner of NFDI4Culture, and the broader family of Wikimedia projects.

Linked open data (LOD) provides many opportunities for retrieval, data enrichment and question answering. However, tools that are built for these purposes are only useful if the data they are using is as complete as possible. Furthermore, there is a lack of established workflows to guide the digital curators, researchers and librarians who manage scientific, cultural or other institutional data through the process of making their data available as LOD, even though there are already tools that can facilitate this process via accessible, and (relatively) user-friendly interfaces, including tools built and maintained by the international community of Wikimedia project contributors and developers.

NFDI4Culture aims to improve the infrastructure for research data in cultural research areas in Germany. One key area of development is the knowledge graph of cultural objects, places and events, which will reuse datasets from multiple, heterogeneous cultural institutions. Most of these datasets are not yet available in linked data format, for example via a SPARQL interface. Many organisations also lack the expertise to transform their data.

The data upload pipeline [1] presented in this paper was developed as a response to the needs of the NFDI4Culture community, but we believe it can resonate with the needs of researchers and data managers across various national and international institutions who meet the same requirements for making their data more easily accessible and reusable.

Wiki Workshop 2022, The Web Conference 2022, April 25, 2022.

Furthermore, this data pipeline interfaces heavily with Wikimedia projects such as Wikidata, Wikibase, and soon also – Wikimedia Commons.

The key tools included in the pipeline are:

- **OpenRefine**: "a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data." [2]. Most crucially for projects intending to use Wikibase and/or Wikidata as a final data repository, OpenRefine allows direct manipulation of data in Wikidata via a reconciliation service and an editing extension. By reconciling values (for example, names of persons, places or cultural objects) to authority control or other linked data services (including VIAF, Getty vocabularies and the GND, among others) OpenRefine also provides options to enrich datasets with new data pulled from these services, thereby reducing manual labour and data redundancy.
- **Wikibase suite**: Wikibase and Wikidata are two related software packages from the Wikimedia family of applications [3]. Wikibase is the open source software environment built to run Wikidata. Wikibase can be deployed independently from Wikidata (and Wikimedia) and can be customized to suit the needs of individual data domains and data repositories. At the same time independent Wikibase instances can use various approaches to federate data and/or data queries with Wikidata. Supporting a federated ecosystem like that is one of the key strategic priorities of the Wikimedia movement [4], as well as various national and international initiatives such as NFDI4Culture and the Wikibase Stakeholder Group [5], respectively.

The pipeline documentation guides users through all the steps needed to clean, transform, reconcile and upload data to an LOD repository, with the long term goal to contribute to the effort of increasing the number of datasets that are part of the semantic web; and also to inspire more participants to collaborate towards the development of this pipeline and the open source tools it depends on. All the tools that are part of the pipeline have graphical user interfaces, lowering the learning barriers to participating in the semantic web, and allowing users with non-technical backgrounds to effectively work with LOD.

CCS CONCEPTS

• Data management systems • Open Source Software

KEYWORDS

OpenRefine, Wikibase, Wikidata, Linked Open Data, FOSS, FAIR data

ACM Reference format:

Lozana Rossenova and Lucia Sohmen. 2022. Opening up Participation in the Semantic Web: A New Data Upload Pipeline. In *Proceedings of The Web Conference 2022, Wiki Workshop 2022, April 25, 2022.*

ACKNOWLEDGMENTS

NFDI4Culture is being funded by the Deutsche Forschungsgemeinschaft (DFG) under grant no. 441958017.

REFERENCES

[1] Lozana Rossenova. and Lucia Sohmen. 2021. OpenRefine to Wikibase: Data Upload Pipeline. Wikiversity. Retrieved 18 Jan 2022 from https://en.wikiversity.org/wiki/OpenRefine_to_Wikibase:_Data_Upload_Pipeline

[2] Retrieved 18 January 2022 from https://openrefine.org/. See also: Elizabeth Sterner. 2019. Cleaning Collections Data Using OpenRefine. In*: Issues in Science and Technology Librarianship.* 92. DOI: 10.29173/istl30

[3] Retrieved 18 January 2022 from https://wikiba.se/ and https://www.wikidata.org. See also: Alípio, S., Abdulai, M.S., Burnett, G. and Shick, D. 2021. Wikibase: the Software for Open Data projects. *Wikimedia Tech News.* Retrieved 18 January 2022 from https://tech-news.wikimedia.de/en/2021/04/14/wikibase-the-software-for-open-data-projects/

[4] Retrieved 18 January 2022 from https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Wikibase

[5] Retrieved 3 February 2022 from https://wbstakeholder.group/about