Project: Extension of Wikibase support in the OpenRefine data cleaning tool, specifically adding support for local media upload in arbitrary Wikibases and adding support for custom data types.

Duration: 6 months (July-December 2022)
Budget: EUR 10,000

Rationale and context

Wikibase is growing in popularity as a tool used by cultural and research institutions to store and structure Linked Data, as well as various media files. It is also written into the NFDI4Culture proposal for the development of a 4Culture Knowledge graph as the preferred tool, benefiting all 4Culture Task Areas.

Managers of distributed, custom Wikibase instances regularly need to perform (sometimes quite sizable) batch uploads and edits of/to their data and media files there. With this grant, the OpenRefine team wants to extend OpenRefine's Wikibase functionalities to support this use case better, by integrating support for local media upload in arbitrary Wikibases and support for custom data types in Wikibases.

[OpenRefine](#) is a free data wrangling tool that can be used to clean tabular data and connect it with knowledge bases. It is a community-supported open source project (licensed under the BSD license). OpenRefine is used by quite diverse communities interested in managing complex data: librarians, researchers, data scientists, and by the broader Wikimedia community, too, who use it to prepare and upload structured data to Wikidata, a free and open public knowledge base. OpenRefine has a graphical user interface which is available in more than 15 languages.

Today, OpenRefine's Wikibase extension already supports batch uploads and edits of/to metadata on Wikidata and arbitrary Wikibases. By June 2022, support for upload and batch edit of files on Wikimedia Commons (Wikimedia's media repository) will also be added to the Wikibase extension ([funded by a grant from the Wikimedia Foundation](#)). Development through this (NFDI) grant will build further upon these existing features.

Deliverables

By the end of 2022, the OpenRefine team will deliver the following:

Batch upload and batch editing of media files in Wikibases through OpenRefine

As a result of new features in OpenRefine developed through this grant, managers of, and contributors to a Wikibase instance will be able to upload large batches (up to 10,000s) of

media files to an arbitrary Wikibase, leveraged by OpenRefine. In addition, they can also edit (modify, add to, delete) the (structured) metadata of the media files stored in their Wikibase, through using OpenRefine.

- In order to make this possible, media file upload functionality through OpenRefine will be modified and made more flexible. With this NFDI grant, we will also add upload and editing support of media files to any Wikibase. This includes adding support for a new data type in OpenRefine (local media file in a Wikibase).

Batch editing of data stored in custom (non-Wikidata) data types in Wikibases through OpenRefine

As a result of new features in OpenRefine developed through this grant, managers of, and contributors to a Wikibase instance can (batch) edit data in any (custom, atypical) data type defined in their Wikibase through using OpenRefine

- In order to make this possible, Wikibase data type support in OpenRefine will be modified and made more versatile/flexible. In early 2022, OpenRefine's Wikibase extension already supports all [data types that are used inside Wikidata](#). However, cases exist where Wikibase managers want to implement / deploy custom data types which differ from the ones used in Wikidata. One example is the local media file datatype (mentioned above). Another example is the [EDTF data type in Wikibase](#), which is more specific than Wikidata's own [Time datatype](#) and which was commissioned and deployed by the Luxembourg Ministry of Culture. With support through this grant, OpenRefine's data type support will be made extensible so that in the future, OpenRefine can support more data types, even ones that are not yet developed.

Dissemination to NFDI stakeholders

By the end of this grant, the OpenRefine team will
- Produce basic end user and developer focused documentation for the above mentioned features
- Present the new features to interested stakeholders inside NFDI and more broadly through a (in-person and/or online) workshop/masterclass. This will be recorded and the recording will be made publicly available.

Who will work on this?

Antonin Delpeuch is OpenRefine's lead developer who is behind the initial Wikidata integration in the tool. In this project he will build the OpenRefine features described above.

Sandra Fauconnier is OpenRefine's project director and works as product manager for the integration of Structured Data on Commons functionalities in OpenRefine. In this project she

will also act as product manager and will lead dissemination activities.

Lozana Rossenova is NFDI4Culture's Wikibase Community manager and will act as liaison between the OpenRefine team and NFDI4Culture stakeholders for this project. With experience in running the Wikibase Stakeholder Group even before joining NFDI4Culture, she has gained an extensive understanding of institutional requirements for Wikibase development and how external tools like OpenRefine can help meet those needs. Her time on this project will be covered within her already contracted hours at TIB and will not use any additional funding from the FlexFunds.