OpenRefine NFDI Tools Forum funding form:

Title of the project:
OpenRefine

Description (500 words):
OpenRefine is a free data wrangling tool that can be used to clean tabular data, reconcile data entities (i.e. identify matching entities across data services) and connect these with external knowledge bases. It is a community-supported open source project (licensed under the BSD license). OpenRefine is used by diverse communities including: 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.

It is also a commonly used tool among the NFDI4Culture community. A workshop dedicated to OpenRefine and its role in working with Linked Open Data, which was organized during the Culture Community Plenary in 2021, received more than double applications for participation than was possible to accommodate. The data submitted by workshop applicants varied greatly in terms of subject matter and data types, hence the need for a user-friendly tool that can accommodate bulk operations on various data types and data scenarios.

In 2021, OpenRefine received a grant from the Wikimedia Foundation to develop new functionalities that allow the upload of media files, alongside structured metadata, to Wikimedia Commons, a vast repository of open license media maintained by the Wikimedia Foundation, and one important repository of choice for open GLAM content. There are currently no tools that enable batch upload of media files and linked data to Commons, and OpenRefine fills a big user needs gap here.

Besides enabling upload of data to Wikimedia Commons, of interest to communities outside Wikimedia, such as NFDI, is also enabling the reconciliation service and upload functionalities provided by OpenRefine, to connect not only to Commons and Wikidata, but also to arbitrary Wikibase instances with their own structured data schemas and independent media file collections. Wikibase is growing in popularity as a tool to be used by cultural and research institutions. It is also used in Task Area 1 as core data enrichment infrastructure, and plays an important role in the Linked Open Data Working Group managed by Task Area 5 of NFDI4Culture. Enabling extended connectivity between OpenRefine and Wikibase is currently being supported by a small Flex Funds Tools grant from NFDI4Culture for the year 2022.

In the course of working on extending OpenRefine's capabilities to connect to Wikimedia Commons and to Wikibase, the OpenRefine team carried out various user testing and user experience consultation sessions. Thanks to these sessions, the team have been able to identify a number of improvements to the reconciliation process that can significantly benefit the overall user experience. These concern: 1) how users interact with the reconciliation dialog window in OpenRefine; 2) how the interface displays reconciliation results from different services, including Wikidata, Wikibase, but also other standard terminology services such as the GND, Getty Vocabularies, VIAF and more; and 3) how users perform data enrichment on their own data via externally linked services. These improvements are out of scope for the grants concluding in 2022 and will benefit from further funding in 2023.

Applicants:
OpenRefine project (CS&S) / Lozana Rossenova (TIB)

Other collaborators:
Sandra Fauconnier, Antonin Delpeuch

Planned working packages (incl. costs and timeframes):
A key (very popular, and unique) Linked Data software feature inside OpenRefine is reconciliation: the process of matching a list of values in a source dataset to an external database (such as Wikidata, a vocabulary resource like GND or the Getty thesauri, or an arbitrary Wikibase). The reconciliation protocol is managed by a W3C community group, which also maintains a census and test bench of the currently more than 50 reconciliation services managed around the world. Maintainers of Wikibases also increasingly deploy their own reconciliation APIs and perform reconciliation for their own and external datasets. Improvement of the usability and clarity of the reconciliation process inside OpenRefine is a popular request from OpenRefine's user community and has also emerged as a priority from user testing in 2022. As part of a renewed NFDI grant, the OpenRefine team wants to tackle the following packages:

1) Redesign and redevelopment of the reconciliation service dialog interface inside OpenRefine, improving the clarity and ease of use of this dialog.
2) Improving how data that has already been reconciled is displayed to users, particularly in cases where multiple external databases are used as reconciliation endpoints and where confusion between data reconciled against e.g. Wikidata, vs a private Wikibase instance is concerned. Improvements of data display will also cover display of media files (e.g. thumbnail previews), which presents different challenges from displaying data types such as semantic entities, numerical values, or plain text.
3) Improving the data enrichment process, which also relies on reconciliation, but involves working with a separate dialog UI. This working package will cover improvements in the user messaging in the data enrichment UI and improvements in how data is displayed to users after enrichment processes are completed - making the provenance of new data instantly recognizable.

Last but not least, all these packages will involve improvements of error messaging to users that may occur as a result of reconciliation or enrichment, prior to users uploading their data and/or media files to their desired end repository.

Timeframe: the above tasks will be executed over a period of 6 months, January-June 2023.

Costs:
Software development €7,000.00
Product management, documentation, various logistics €1,500.0
Project leadership, administration, accounting, strategic support (Code for Science and Society, OpenRefine's fiscal sponsor) - 15% of budget total - €1,500.00
Total: €10,000.00

Relevant for the following disciplines / target groups:
Because this project concerns metadata and media file management at a high level, it is applicable to a wide range of disciplines from library and archive science, to research with all types of cultural heritage data. OpenRefine has been cited as a tool commonly used by the NFDI4Cuture community at previous events (*https://nfdi4culture.de/news-events/events/byod-workshop-bring-your-own-dataset-how-to-use-wikibase-openrefine-and-linked-data.html*), and is an important tool for the Wikimedia community as a whole.

State of the project:
Further development

What preparatory work exists?

This work is a follow-up to the 2022 work on Wikibase and Wikimedia Commons integration. During those projects, user testing was conducted and identified the importance of those improvements. Design wireframes for the improvements already exist and have been discussed with OpenRefine's community of contributors.

This work would also build on the current effort to improve the reconciliation API, which is done by a group of stakeholders under the auspices of the W3C Entity Reconciliation Community Group. Although this proposed work is motivated by Wikibase integration, it would therefore improve workflows for other data sources: GND, Getty, VIAF and many other reconciliation services.

What user needs will be solved with this project? What goals are pursued?
This project improves two key Linked Open Data functionalities in OpenRefine: data reconciliation and data enrichment, both of these are connected closely with how OpenRefine fits into workflows involving external standard data authority control services as well as workflows involving Wikidata, Wikibase and Wikimedia Commons. The goals of the working packages proposed with this application are to directly address specific user requirements discovered during user testing and consultation sessions carried out in 2022. These requirements concern the user interface design of dialog windows and data display functionalities, which will improve user's understanding of data provenance and speed up workflows thanks to introducing more clarity into the processes of reconciliation and enrichment. Last but not least, better error messaging will make it easier for users to identify precisely what may be going wrong with their data, address the error and continue with their follow-up data tasks, such as data upload to Wikibase, for example.

Are there comparable tools? Could parallel developments be closed?
Currently, no comparable tools can perform the complete workflows possible with OpenRefine – from data cleaning, to reconciliation, enrichment and finally data upload to a public or private Wikibase instance. For parts of these workflows, scripts can be used, but this requires programming skills and individual adjustment for every dataset. This work will not constitute parallel development to anything else under development in NFDI4Culture. On

the contrary, it will support and enhance ongoing work on Wikibase development within Task Areas 1 and 5.

What are the unique selling points of the project?
OpenRefine is the only tool that can complete an end-to-end user workflow from data cleaning to data upload in a Linked Open Data repository with the help of a graphical user interface. What is more, the interface is available in more than 15 languages, including German – making this a highly accessible tool. The user-friendly interface is particularly important for the culture community, where many humanities researchers are experts in their data, but lack the programming skills to develop their own data wrangling scripts.

Importantly, OpenRefine is a free and open source tool, which can be built upon and developed further by the culture community as required. It is supported by a global community of open source developers active on the project's Github repositories.

What is the impact of the project on the community?
The improvements to the reconciliation and data extension user workflows proposed in the working packages of this funding application will significantly improve the user experience of the tool as a whole, and specifically the connectivity to external terminology services and data repositories which makes the tool highly valuable for the culture community. Since these improvements were developed in the course of user testing with actual culture community members, we believe these are essential and will have a proven positive impact on how users work with their data in OpenRefine.

To what degree can the sustainability of the project be guaranteed?
OpenRefine is an open source project with a healthy, active, international community and governance structure. OpenRefine actively maintains its product roadmap and is supported by a fiscal sponsor (Code for Science and Society) which aids the project in its administration and strategy. The software has been actively used and developed for over 10 years, and is popular with end users from a variety of backgrounds: LOD specialists, cultural heritage specialists, data scientists, journalists, Wikimedians and Wikibase administrators. Besides NFDI, OpenRefine receives funding from other sources: donations, and grants from organizations like the Wikimedia Foundation and the Chan Zuckerberg Initiative. This broad community support and diversity of income ensures continued support for new features that receive adoption from end users.