NGi Zero Commons Fund application: Reconciliation for Wikibase

Application form: https://nlnet.nl/propose/
Person submitting the application: Antonin?
Application deadline: October 1st, 12:00 CEST

Submitted!

Abstract: Can you explain the whole project and its expected outcome(s). (max 1200 characters)

When contributing to or reusing knowledge commons such as Wikidata or OpenStreetMap, we often need to match external datasets to existing entities inside the collaborative database. This matching - which we also call reconciliation - makes it possible to enrich the knowledge common from the external dataset, or conversely.

The reconciliation API is a web-based protocol which eases this process. The protocol has been made popular by OpenRefine (an open source data cleaning tool), its historical client.

We want to improve the support for the reconciliation protocol in Wikibase, which powers Wikidata (the leading collaborative knowledge graph) but also the EU knowledge graph, the OpenStreetMap wiki, and many more. So far, reconciling data with a Wikibase is only possible by running a wrapper web service on top of Wikibase. This Python service has been poorly maintained for multiple years, primarily because its wrapper architecture makes it hard to offer a satisfactory experience for end users.

We propose to develop a Wikibase extension implementing the reconciliation protocol natively. We expect this will greatly ease the deployment and use of this functionality.

Have you been involved with projects or organisations relevant to this project before? And if so, can you tell us a bit about your contributions?

I am the original author of the wrapper web service to be replaced (written in 2017). I am also an OpenRefine developer, where I worked on improvements to reconciliation over the past few years, either directly or as an Outreachy mentor. Finally, I co-chair the W3C Community Group about this protocol, which has been formed in 2019 to steer the evolution of the API.

I represent the OpenRefine project in the Wikibase Stakeholder Group (WBSG), a network of Wikibase users which will coordinate the development of this Wikibase extension.
The implementation will be done by Professional.Wiki, a team with considerable experience developing Wikibase extensions. I will assist with project coordination, share knowledge acquired from developing my wrapper service and from maintaining OpenRefine.

Requested amount

50,000€ (maximum)

Explain what the requested budget will be used for?

Does the project have other funding sources, both past and present?
(If you want, you can in addition attach a budget at the bottom of the form)

The expected costs are as follows:

MVP development: 45 days, 31.500€

- Wikibase extension with new API endpoints for core reconciliation capabilities
- Extension internationalisation and automated testing
- Management and product work
- Documentation, including README, blog post, and YouTube video with demo

Stretch goal A: Additional reconciliation capabilities, 35 days, 24.500€

- Scoring algorithm improvements
- Data Extension Query Requests endpoint
- Entity preview HTML endpoint

Stretch goal B: Further user experience enhancements, 20 days, 14.000€

- Support for Properties and other entity types
- In-wiki configuration UI (usable on cloud hosting and elsewhere)

The cost estimates are based on an average day rate for senior frontend/backend engineer based on this Germany 2024 statistics:
https://www.malt.de/t/tarifbarometer/it/backend-entwickler?

We are able to break down the goals above into more granular work packages.

In addition to the NGi Zero Commons Fund, we are planning to fund work on this project from multiple sources:

- The Department of Culture Flanders and from Kunstenpunt pledged 7.500€
- We applied for funding from the NFDI4Culture Tools forum (expecting to contribute between 5.000€ and 10.000€)
- We are hoping to get a contribution from Virginia Tech (amount to be determined)

Compare your own project with existing or historical efforts.

We are aware of various existing wrappers offering reconciliation on top of Wikibase:

- The wrapper I wrote, running at https://wikidata.reconci.link/, not actively maintained
- A fork of it with improved documentation on deploying it to third-party Wikibases: https://github.com/judaicadh/wikibaseopenrefine
- An older implementation for Wikidata by Magnus Manske, which is no longer operational: https://toolhub.wikimedia.org/tools/mm_wd_reconcile
- A wrapper that works for any Wikibase instance, without requiring configuration, but not offering the full range of features allowed by the protocol, and proprietary: https://byabbe.se/2024/08/05/reconcile-against-any-mediawiki-instance

This proliferation of wrappers demonstrates the need for a native integration and the frustrations around existing solutions.

The Wikibase Stakeholder Group has developed various extensions for Wikibase: https://wbstakeholder.group/projects/extensions
Some of those have been included in the official Wikibase Suite product maintained by Wikimedia Deutschland, which is a sign of their popularity.

What are significant technical challenges you expect to solve during the project, if any?)

The implementation will be contracted out to Professional Wiki, who are experienced in developing Wikibase extensions. We have identified the following challenges:

End users are used to the behaviour of the current wrappers. For instance, its scoring mechanism is far from perfect, but it has been stable for many years. With this extension, we have the opportunity to offer a better score given that we will have closer access to the search indices, but replicating the exact behaviour of the current scoring might be hard, as it relies on a particular Python library for fuzzy-matching.

Another challenge will be the implementation of the type filtering. The current wrapper relies on the query service to fetch the collection of subtypes of a given type, which is a significant performance issue. We should ideally avoid this, perhaps taking inspiration from how type checking is done in the WikibaseQualityConstraints extension (which relies on a dedicated query service for more reliable performance).

Describe the ecosystem of the project, and how you will engage with relevant actors and promote the outcomes?

The relevant groups for this work are:

- Wikibase end users, in particular those who use OpenRefine with their Wikibase (or would like to)
- Wikibase administrators (gathered in the Wikibase Stakeholder Group, WBSG), who want to deploy reconciliation functionality easily

- The W3C Community Group, which can be approached to suggest changes to the protocol if the need arises
- The OpenRefine project, as canonical client for the reconciliation API (also WBSG member)
- Wikimedia Deutschland, as maintainers of the Wikibase Suite (also WBSG member)

The meetings of the WBSG have been serving as a regular coordination gathering for this project. We plan to continue using this network to reach out to Wikibase users and administrators to gather their feedback throughout the project.