Hello, what's the name of your initiative?

OpenRefine: Improving OpenRefine's Extension System

In what country will your initiative have impact?

Worldwide.

Who is your primary contact?

This is the person and organisation that would lead delivery of the initiative, and would be the recipient of funding from the Data Empowerment Fund.

First name: Martin
Last name: Magdinier
Email: martin@openrefine.org
Company: Code for Science and Society, Inc.

Describe the problem you are working to solve, or your diagnosis of the status quo and why it must change.

OpenRefine is a popular tool for communities collaborating on a common knowledge graph. Crowdsourcing data from a grassroots group is challenging, as the contributors need to express their contributions in a standard format and link their additions to existing records to avoid duplication.

OpenRefine offers a plug-in system, enabling third-party teams to develop integration for other crowdsourcing platforms and publish them as OpenRefine extensions. Those extensions enable users to match, link, and publish their data to an existing dataset. We can cite:

- The two RDF Extensions (grefine-rdf-extension and rdf-transform) are our second most installed extensions. They enable the connection between OpenRefine and SPARQL or RDF endpoints. The example below presents different use cases where academics used OpenRefine and the RDF extension to build new datasets
-
    - Developing a Knowledge Organization System for Ethnic Groups in Lao PDR through Linked Open Data Techniques
    - A knowledge graph of interlinking digital records: the case of the 1997 Korean financial crisis

    - Developing a Linked Open Data Platform for Folktales in the Greater Mekong Subregion
- GeoJSON Export, which generates GeoJSON format,
- SNAC Extension for publishing to the Social Networks and Archival Contexts knowledge graph,
- FAIR metadata for storing and exporting data to FAIR format
- CKAN storage for connecting OpenRefine with CKAN, which facilitates the publication of open data set by governments

In 2018, the OpenRefine team developed its own integration with Wikidata. It has been widely adopted by the Wikidata community as an essential data import tool, without requiring users to have programming skills. It has since been generalized to work with other Wikibases (in 2020) and with Wikimedia Commons (in 2022).

However, those integrations are tailored to specific models, making it useless for other communities working with different platforms. When developing those extensions ourselves, we realized that OpenRefine's integration process is hard to use both for end users and for extension developers.

For users, discovering and installing an extension is cumbersome and error-prone. Upgrading extensions or OpenRefine often breaks functionality because of incompatibilities between versions.

For developers, OpenRefine's extension mechanisms are unclear, making plugins complicated to maintain in the long term. As a result, most of the extensions developed in the last years are incompatible with OpenRefine's latest version.

Describe how your initiative is designed to address this problem or bring about this change.

Please make this short and sweet. We will ask for information about specific activities, outputs and outcomes in a moment.

We want to improve OpenRefine's extension system so that it can be used reliably by third parties to build the features their community needs. We also want to develop an extension manager to make it easier for users to discover, install, and manage extensions.

Tell us about the composition, experience and diversity of your team.

Also list any other organisations who would be involved in your initiative, such as delivery partners or advisors.

Core team:

- Antonin Delpeuch (developer), main author of the Wikidata integration in OpenRefine, joined the project in 2017. Based in Leipzig, Germany.
- Zoe Cooper (designer), with a background in journalism, joined the project in 2023. Based in Berlin, Germany.
- Martin Magdinier (project manager), experienced OpenRefine instructor and community organizer, joined the project in 2012. Based in Montreal, Quebec

Partners:

- Wikimedia Sweden, who have taken over the maintenance of the Wikimedia Commons integration in OpenRefine.
- Other maintainers of existing extensions (RDF Transform, OpenRefine Command Palette, SNAC extension) and Organizations which have chosen to fork OpenRefine rather than build extensions for it (OntoText, Fornpunkt.se. We want to reach out to them to understand how we can simplify our extension system.
- NFDI4Culture and the Wikibase Community User Group, with which we previously funded work to improve the OpenRefine reconciliation feature.

Advisory Board:

- Antonin Delpeuch
- Jan Ainali
- Esther Jackson
- Julie Faure-Lacroix

Describe the maturity of your initiative.

Tell us how long it's been running, if you have participants involved already and any impact it has achieved to date (if any).

OpenRefine is an open source project which started in 2010. It has been fiscally hosted by Code for Science and Society since 2019. To date, 328 contributors have made changes to its source code, translations or documentation. It is regularly taught in training workshops around the world that are addressed to diverse user communities such as journalists, librarians, Wikimedians and open source intelligence practitioners. On average, OpenRefine is downloaded 15,500 times per month and receives over 300 academic citations per year.

We are aware of at least 25 publicly available OpenRefine extensions, although many of those have not been updated to work with recent versions of the tool.

What value of grant are you seeking from the Data Empowerment Fund?

USD 100,000

Mini-budget for a USD 100,000 grant not asked in the grant application:

- CS&S 15% 15,000$
- 2 contractors
  - Designer 1 month or 2 months 15,000$
  - Developer part-time for 9 months: 50,000$
- 3 to 5 grants to update existing extensions, including OpenRefine project manager time USD 20,000$

Describe what activities you would use the grant to undertake.

We have decided to split the project into two separate streams of activities that will happen concurrently.
Stream 1: Improve OpenRefine's support of extensions:

1. Let users install extensions directly within OpenRefine itself using an extension manager.
2. Improve our documentation for extension developers. A Google Summer of Code intern may carry out initial work on this front, as we have proposed it as a possible internship subject (https://github.com/OpenRefine/OpenRefine/wiki/GSoC-Outreachy-2024-Ideas#improve-the-extension-development-process).
3. Address elements we want to improve for extension developer (see https://forum.openrefine.org/t/improving-the-ux-of-extension-install-and-butterfly/52 and https://github.com/OpenRefine/OpenRefine/labels/extension).

We plan to organize the stream 1 based on the following timeline.
Months 1 to 3: team building

- Hiring the developer for the implementation, as our lead developer is already fully committed to another project until Dec 2024.
- We already have a designer on staff.

Months 4 to 6:

- Design of the interface of the extension manager
- Onboarding of the developer on the code base
- Validate the scope of the changes done to the integration and impact for the migration with current extension developers.

Months 6 to 12:

- Implementation of the design and development.
- Improvements to the current documentation and tutorials.

Stream 2: Provide mini-grants to help maintain existing extensions.

Months 1 and 2: Create the scope of the call for proposals

Months 3 and 4: Open the call for proposals to upgrade and maintain existing extensions following breaking changes introduced in previous versions of OpenRefine and improve their test coverage. We will individually reach out to each extension maintainer to promote the call for proposals.

Month 5: Select proposal(s)

Months 6 to 12: Work on extensions with the support of the developer

Describe what outputs you would use the grant to produce.

These are the things your activities will produce.

The grant from the Data Empowerment Fund will be used to produce the following outputs for the OpenRefine community:

Integrated extension manager accessible directly in OpenRefine, making it possible to install an extension directly from the tool interface.

Improved documentation and tutorials to assist developers in creating extensions and maintaining them in the long term.

Updates at least three Existing Extensions: so they are compatible with OpenRefine's latest version.

Describe what outcomes you would use the grant to achieve.

These are the changes in behaviour, decisions or knowledge that your activities or outputs will bring about. Tell us about the people or communities that will most benefit.

With those improvements, we'll first and foremost make our existing users' life easier as they will be able to install and remove extensions more easily. We'll also make extension developers' life easier, by reducing the maintenance burden of their extensions. On a social level, investing in the maintenance of their extensions will show them that we value their work and want to support them, increasing cohesion in the ecosystem around OpenRefine.

Ultimately, we want to provide more users with individual agency to their local knowledge graphs and control their community data, thanks to the upgraded extensions.

Describe the ultimate impact you would use the grant to achieve.

This is the longer term result of your work, likely to occur beyond the grant period.

With a more mature extension system, we hope that more communities will decide to develop integrations of their platform for OpenRefine, allowing us to support more data crowdsourcing projects. This will make it easier for users without extensive technical backgrounds to make sizable contributions to those projects. We have already seen success with Wikimedia projects thanks to our existing extensions.

We also hope that some of the features currently developed within OpenRefine could be migrated to extensions. This would make OpenRefine less monolithic, making its maintenance easier both from a technical and social perspective, as it would become easier to have different people responsible for different areas of work.

What will others be able to learn from your initiative?

We're particularly interested in supporting initiatives that: have significant public policy relevance; **enable people to control how data is used to train AI models**; and/or involve a novel legal, technological or **participatory** approach.

We hope we can serve as an example for scaling a particular open-source tool to satisfy the crowdsourcing needs of distinct communities, particularly those that work with data that may be used to train large AI language models. By enabling a healthy extension ecosystem, we hope to enable those communities to take ownership of the tooling they rely on, and give them greater agency over their data and its reuse.

Describe how the Data Empowerment Fund could support your work beyond providing access to funding.

E.g. by providing guidance on a particular challenge you're facing, or by connecting you to particular types of expertise.

We would like to get in touch with other communities with data integration needs, such as other crowdsourced knowledge bases, potentially leading to new collaborations.

We are currently in the process of improving our governance model and are interested in discussions with similar projects mixing grassroots contributions with paid roles.

Are you happy for us to share your application with our partners, the Patrick J. McGovern Foundation and Omidyar Network?

If your application is unsuccessful to the Data Empowerment Fund, it may be a good fit with their other interests.

Yes