Proposal Title

Supporting Biomedical communities in using OpenRefine

Amount Requested

We request USD 400k total, split USD 200k per year for two years.

Proposal Summary - Scope of Work (500 words)

OpenRefine is a free, open-source software offering advanced data quality and cleansing features, including data normalization, duplicate removal, pivoting, joining, enrichment through third parties via API, and data splitting.

Thanks to the EOSS-5 and Diversity Grant from CZI, we gained valuable insights into the pivotal role trainers play in OpenRefine's ecosystem. Through our survey, community forum, and direct outreach, we realized that trainers serve as ambassadors, advocates, and educators. Despite their importance, they often operate independently, lacking adequate support. The EOSS-6 grant seeks to formalize and enhance OpenRefine's backing for trainers. This initiative ensures better support for the biomedical community and strengthens the connection between OpenRefine and its diverse user base.

1. Community Engagement and Training Support

We have identified the following options to support trainers. Before proceeding, we will conduct a thorough needs assessment within the biomedical community to ensure the relevance and effectiveness of each activity.

Firstly, by hiring a community manager, we want to establish a formal communication channel between the trainers and the OpenRefine developer. This will enable the developers to keep trainers informed about new features and upcoming releases, while trainers can provide valuable feedback on software usage, feature requests, and help guide the roadmap.

We want to offer direct financial support to existing trainers and empower new ones. Potential activities include organizing training, webinars, creating tutorials, translating OpenRefine documentation, and allowing them to answer support requests from their community. By tying funding to specific deliverables, we aim to produce tangible outcomes that closely align with the OpenRefine long-term strategy. We plan to initiate an open call for applications, leveraging existing community channels and relationships with open science organizations like the Research Software Alliance.

Finally, building on insights gathered during our Diversity grant outreach, we want to improve the OpenRefine user onboarding experience. We plan to provide an easy, self-guided interface

tour and improve the UX to simplify new user orientation and reduce the efforts required by trainers.

2. Technical Position for GitHub Repository Maintenance

The proposal includes the renewal of the role of the OpenRefine lead developer. Over the past four years, OpenRefine 4.x version has undergone significant enhancements, made possible through the support of the EOSS-1 and EOSS-5 grants. These enhancements include critical features for the biomedical community, such as support for larger datasets and streamlined reproducibility and workflow automation.

Looking to the 2025-2026 period, the lead developer will focus on ensuring the stability and continuous improvement of the new version, especially as those new features are incorporated into the workflows of biomedical researchers.

The lead developer's responsibilities include key activities we previously identified as essential in our previous grant application to maintain sustainable and diverse contributors communities. These tasks include maintaining the OpenRefine GitHub repository, triaging tickets, reviewing pull requests, participating in the Outreachy program, and onboarding new technical contributors. Additionally, the lead developer will implement new features and address bugs reported by the trainer community.

Value for Biomedical user (250 words)

OpenRefine serves as an essential tool for biomedical researchers, supported by insights from our 2022 user survey involving 50 participants with a research background (out of a total of 178 answers) and relevant literature reviews focusing on publication in 2022 and 2023 only.

First 75% of respondents use OpenRefine at least once per month, with an equivalent proportion having used the tool for more than a year. This underscores OpenRefine's integral role in the workflows of researchers.

Researchers employ OpenRefine for diverse tasks, with 94% utilizing it for cleaning and normalizing data—a critical role in maintaining data quality. This is evidenced in publications such as:

- Patients' Severity States Classification based on Electronic Health Record (EHR) Data.
- Assessing spatial and temporal patterns and risk factors for tick acquisition.
- Analyzing Autism Prevalence Among Original Medicare Beneficiaries.

Additionally, 60% of users employ OpenRefine for preparing data for import into other systems and working on taxonomies and thesauri as seen in publications such as:

- An Ontology-Based System for Cancer Registry Data.
- Semantic resource responding to Open Science principles: The meat thesaurus.

Half of the respondents use OpenRefine for reconciliation against other datasets, a speeding up thesaurus and bibliometric analyses, evident in publications including:

- Antifungals from plants: a bibliometric analysis
- Potential use of microalga Dunaliella salina for industrial bioproducts
- Modeling COVID-19 Transmission Dynamics: A Bibliometric Review.
- Coral restoration patents and academic research disconnection.
- Assessments of two-pore channel 2 in the human MDAMB-231 breast cancer cell line.

Open Source Software Project

Software project name	**Main code repository (e.g. GitHub URL)**	**Homepage URL (if none, re-enter the main code repository URL)**
OpenRefine	https://github.com/OpenRefine	https://openrefine.org

Landscape Analysis (250 words)

Data cleansing tools can be categorized as follows:

1. Spreadsheet software provides an entry-level interface to data manipulation, but offers only basic functionalities and does not scale for the large datasets commonly used in science contexts.
2. Programming languages like Python and R offer flexibility and reproducibility but have a steep learning curve.
3. Data workbench like KNIME or SAS data analytics, reporting, and integration using visual components
4. Data preparation software like OpenRefine fills the gap between these categories. With a powerful GUI, this category of software can be easily mastered by non-programmers, also supporting large datasets.

Free and open-source data manipulation software makes this functionality available to communities and user groups with few resources as well. Solutions include:

- Orange http://orange.biolab.si/ - focus on data visualization
- Workbench https://github.com/CJWorkbench - focus on data journalism; ceased operations in 2021
- QSV https://qsv.dathere.com/ released in 2023 currently offer limited functionality.

OpenRefine stands out in the data preparation category, bridging gaps between spreadsheet

software and programming languages. With its robust GUI, OpenRefine is designed for non-programmers while supporting the handling of large datasets. OpenRefine is the most advanced interface for data reconciliation and linkage within the biomedical research context. Overall OpenRefine provides a low-code interface, making data exploration, cleaning, and preprocessing tasks user-friendly.

Category

Data Management and workflow