Title

Enter the title of your application. Proposal title is limited to 60 characters, including spaces. If you need to edit your proposal title, click on the My Applications link in upper right and click the green Continue button on the application you wish to edit. Once the application page opens, click on the three dots to the right of the application title and select Rename from the dropdown menu.

Improving OpenRefine reproducibility

Task 1: Applicant Details.

The information entered should be for the individual submitting the application who will act as the main person responsible for the application and as its point of contact.

Name and email are auto-filled. To edit your name or email, please do so in your account information by clicking your name in the upper right corner and clicking My Account in the dropdown menu.

Sandra Fauconnier sandra@openrefine.org

Institution/Affiliation:

Add your home institution, company, or organization. This does not need to be the organization to which a grant would ultimately be awarded, if selected for funding.

Code for Science & Society (OpenRefine's fiscal sponsor)

Task 2: Proposal Details.

All sections are required.

Proposal Title:

Auto-filled; Maximum of 60 characters, including spaces. If you need to edit your proposal title, navigate to your application summary page; click on the three dots to the right of the application title; and select Rename from the dropdown menu.

Amount Requested:

Total budget amount requested in USD, including indirect costs; this number should be between $100,000 USD and $400,000 USD total costs over a two-year period.

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

Proposal Summary/Scope of Work:

Provide a short summary of the work being proposed (maximum of 500 words).

OpenRefine is a generalist data manipulation tool which serves the needs of diverse communities: scientists and researchers (including the biomedical field), but also data journalists, Linked Open Data practitioners, librarians and cultural heritage specialists, and Wikimedians. OpenRefine offers advanced data quality and cleansing features, including data normalization, duplicate removal, pivoting, joining, enrichment using third parties via API and splitting data.

In the past years, OpenRefine has invested in the growth of its communities and has improved its architecture to support larger datasets. As a next step, in OpenRefine's two-yearly user surveys, scientific communities express the need for better reproducibility and automation of workflows, which the OpenRefine team hopes to develop with the support of an EOSS Cycle 5 grant.

Improvements to OpenRefine's architecture: reproducibility and automation of workflows

We want to improve the verifiability and reproducibility of scientific research done with the help of OpenRefine, by enabling more legible, more flexible and persistently shareable and publishable workflow exports. It is already possible to extract the history of OpenRefine projects as JSON objects and to reapply these to other projects; this is a flagship feature of OpenRefine. As an example from the biomedical domain, a custom solution has been built to automate wrangling COVID-19 related data from the John Hopkins University's COVID-19 repository: https://github.com/dathere/covid19-time-series-utilities.

However, OpenRefine's support for repeatable workflow is partial. In order to make this feature viable we would need to improve workflow visualization, edition and publication, while officially supporting headless execution of scripts. The OpenRefine community has also expressed the need to automate and repeat batch data operations as part of a wider pipeline, and we would like to better support this.

Antonin Delpeuch conceptualized the idea in his publication A Complete Language for Faceted Dataflow Programs (https://arxiv.org/pdf/1906.05937.pdf) and the feature has been extensively discussed on OpenRefine's mailing list:
https://groups.google.com/g/openrefine-dev/c/42mdP8gyt4M/m/s21fJ3W6BQAJ.
We also see community efforts to make this happen, for instance in the project in which OpenRefine histories are visualized as YesWorkflow diagrams:
https://www.ideals.illinois.edu/handle/2142/109699 or an extension to reuse OpenRefine in automated processes https://github.com/openculturecons ulting/openrefine-client

We want to learn from and generalize such custom approaches, making this functionality available as a standard feature in OpenRefine's ecosystem.

Community support and design research focused on the biomedical and broader research communities

In 2020-21, grants for OpenRefine have mostly focused its developers' attention on use cases from the Linked Open Data and Wikimedia communities (including Wikibase). With the help of a renewed EOSS grant, OpenRefine will be able to again focus its attention to its scientific user community as well. For this purpose, we aim to invest in community management, design research and dedicated UX design specifically focused on (biomedical) scientific use cases, and to grow the engagement of biomedical research communities in OpenRefine's governance.

Value to Biomedical Users:

Describe the expected value of the proposed work to the biomedical research community (maximum of 250 words).

Data cleaning and preparation is a significant hurdle for biomedical research, yet access to clean and reliable data is the cornerstone for any analytics and scientific project. For nearly ten years, OpenRefine has served the needs of data science communities.

In biomedical research alone, OpenRefine is used in research projects related to genomics, Alzheimer's disease, infectious diseases, oncology, and clinical data management. Typically, OpenRefine is used for cleaning and manipulating external datasets (for instance related to medical trials and drug reports), and for bibliographic analysis. Since 2020, OpenRefine was mentioned in over 200 academic papers related to COVID-19. Examples of recent biomedicine-related publications mentioning OpenRefine include

- Data Quality usage: Hidden in our pockets: building of a DNA barcode library unveils the first record of Myotis alcathoe for Portugal https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7403162/
- Data processing and enrichment: Open Access of COVID-19-related publications in the first quarter of 2020: a preliminary study based in PubMed https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7438966.2/
- Data cleansing: Pathways in the Drug Development for Alzheimer's Disease (1906-2016): A Bibliometric Study https://jscires.org/sites/default/files/JScientometRes-9-3-277.pdf
- RDF Generation and Data FAIRification: A catalogue of 863 Rett-syndrome-causing *MECP2* mutations and lessons learned from data integration https://www.nature.com/articles/s41597-020-00794-7

Open Source Software Projects:

Indicate the number of software projects involved in your proposal (up to five). Complete the table with the following information for each 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:

Briefly describe the other software tools (either proprietary or open source) that the audience for this proposal primarily uses. How do the software project(s) in this proposal compare to these other tools in terms of user base size, usage, and maturity? How do existing tools and the project(s) in this proposal interact? (maximum of 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 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.

Proprietary solutions include:

- Trifacta/Paxata https://www.trifacta.com/
- RapidMiner https://rapidminer.com/
- Rattle https://cran.r-project.org/bin/windows/base/
- KNIME https://www.knime.org/knime-analytics-platform
- H2O http://www.h2o.ai/download/h2o/choose
- Alteryx https://www.alteryx.com/

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
- Data Preparator http://www.datapreparator.com/downloads.html - low maturity
- Tanagra http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html - low maturity
- Workbench https://github.com/CJWorkbench - focus on data journalism; ceased operations in 2021

OpenRefine is one of the most mature projects in this domain. The project has seen wide adoption for over 10 years; its user base, community and functionalities have been steadily growing. Increasingly, OpenRefine's interface is translated in over 30 languages and (with support from a 2021 EOSS Diversity grant) the tool is being adapted to more linguistically and culturally diverse use cases.

Category:

Choose the two categories that best describe the software project(s) audience:

- Bioinformatics
- Single-cell biology
- Structural biology
- Clinical research
- Genomics
- Neuroscience
- Infectious disease
- Imaging
- **Data management and workflows**
- **Machine learning and data analysis**
- Visualization

Previous CZI funding:

Did you previously apply for funding for this or a related proposal under the CZI EOSS program? Select Yes or No.

Yes

Have you previously received funding for this proposal under the CZI EOSS program? Select Yes or No.

Yes