EOSS5 Full Application

Application

Proposal Details

Please complete the following proposal information. All sections are required.

1. Proposal Title: Improving OpenRefine reproducibility

2. Previous Funding

Did you previously apply for funding for this or a related proposal under the CZI EOSS program?

- No
- **Yes**

Have you previously received funding for this proposal under the CZI EOSS program?

- **No**
- Yes- please specify application number, e.g. EOSS-0000005345, EOSS2-0000005145

3. Proposal Purpose:

Describe the purpose of the proposal in one sentence (maximum of 200 characters including spaces). Example: To develop a comprehensive, validated atlas of the human kidney at single-cell resolution open to the entire scientific and clinical community.

Improve OpenRefine to empower scholars without programming experience to publish research datasets along with verifiable and reproducible workflows, and to automate such workflows.

4. Amount Requested:

Enter the amount requested per year between $50,000 USD and $200,000 USD per year, including indirect costs), as well as the total budget requested for all years in USD (between $100,000 USD and $400,000 USD total, including indirect costs). These numbers should match those described in the Budget Description to follow. Enter whole numbers only (no dollar signs, commas, or cents)

Year One: 155,050
Year Two: 155,050
Total All Years: 310,100

5. Proposal Summary:

Provide a short summary of the application (maximum of 500 words) (auto-filled from LOI; update if needed)

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.

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.
However, OpenRefine's support for repeatable workflows 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. 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.

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 visualised as YesWorkflow diagrams: https://www.ideals.illinois.edu/handle/2142/109699 or an extension to reuse OpenRefine in automated processes: https://github.com/opencultureconsulting/openrefine-client
We want to learn from and generalise 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.

6. Work Plan:

A description of the proposed work for which funding is being requested, including resources the applicants will provide that are not part of the requested funding. For software development-related work (e.g., engineering, product design, user research), specify how the work fits into the existing software project roadmap. For community outreach related activities (e.g., sprints, training), specify how these activities will be organized, the target audience, and expected outcomes (maximum of 750 words)

Improving the reproducibility and automation of data operations is a major long-term request in OpenRefine's roadmap (https://github.com/OpenRefine/OpenRefine/wiki/Roadmap).

After a set-up period (hiring and onboarding contractors for UX design and community management), this work will be tackled in four phases:

1 – User research (November 2022-February 2023)

We will analyze typical data cleaning projects in research contexts, with an emphasis on the biomedical domain. As reproducibility matters for other communities as well (such as general data science and data journalism), these will be surveyed too. We will interview users (focusing on user needs and user stories), while coordinating with the development team (focusing on ideation and technical feasibility). Finally, we will compare the approaches to reproducibility in other tools and languages (such as the Common Workflow Language, FAIR Principles, Protocols.io, Extract-Transform-Load tools, and scripting languages such as Python and R).

2 – UI/UX design and technical specifications (March-April 2023)

Given the outcomes of the user research phase, we will propose changes in the tool to address the most pressing needs. Designs will be proposed through wireframes and technical specifications. Naturally, the research informs the development needs for those new features and plans will be adjusted accordingly.

3 – Software development (May 2023-August 2024)

This is a natural follow-up to our work on scalability (funded by our EOSS1 grant), during which Antonin Delpeuch improved the architecture of the tool with the workflow automation project in mind. Antonin Delpeuch remains our technical lead for this project..

Depending on the outcomes of user research, development activities may at least include:

- Introduction of new methods for the visualization of transformation workflows in OpenRefine, representing the dependencies between the operations applied. This will require significant user interface design and testing to ensure that the visualization fits a wide range of use cases, including the publication of research datasets along with their transformation workflows in a transparent, verifiable and reproducible manner.
- Better error handling when applying workflows with unmet dependencies. The current functionality has deficient error handling which can result in undefined behavior, with some operations being silently skipped. We want to make errors when executing a workflow more visible and intelligible.
- Inclusion of importer and exporter settings in machine-readable representations of workflows. Currently, the history tab of the tool only exposes metadata for the operations, but not the importer and exporter used. To run a workflow (for instance as part of a pipeline), one needs to also represent those.
- Ability to edit workflows, such as reordering operations or changing the parameters of an earlier transformation. When cleaning data interactively in OpenRefine, it can happen that one realizes that a particular operation should have been run before another, or that an operation should have been run with different facet settings. We can combine this with the new workflow visualization feature if it is developed.
- In order to make the automation of workflows possible, OpenRefine needs native command line support. There already exist some command-line tools to run OpenRefine workflows on a dataset without going through OpenRefine's web based UI. Those tools run OpenRefine's backend in the background and communicate with it via our internal HTTP API, which comes without any documentation or stability guarantees. A newly built tool could instead run the workflows directly, without running an HTTP server. Furthermore, for some workflows, this tool could run the operations in a streaming fashion, reducing memory requirements and speeding up the process.

Development will be carried out iteratively, regularly releasing prototypes which will be tested by users who participated in phase 1. The developer community will be consulted about the introduction of versioning and stability policies for OpenRefine's workflow elements (importers, operations, facets, exporters) and expression languages (GREL), which is necessary to guarantee reproducibility in OpenRefine workflows.

A community liaison acquainted with the use of scientific datasets in the biomedical domain will act as ambassador for this project, organizing a range of activities (conference presentations, training sessions) that invite feedback and keep the biomedical research community informed. In addition, we intend to recruit a member from this community into either OpenRefine's advisory or steering committee.

4 – Documentation and training (finalized May-October 2024)

We will add dedicated end user documentation along with training materials to OpenRefine's existing documentation (https://docs.openrefine.org). In addition, a dedicated update to OpenRefine's technical documentation will help OpenRefine's sustainability, easing the onboarding of new contributors.

7. Milestones and Deliverables:

Milestone: One month into the project, the OpenRefine team is supplemented with one or more contractors combining UX/UI design and community management skills who will support this project.

Milestone: At the end of phase 1 (by month 4), initial user research concludes; by that time, OpenRefine has consulted at least 8 scholars and potential end users of the reproducibility features to be developed through this project.

- Deliverable: Publication of a public research report summarizing the findings of the initial user research phase.

Milestone: At the end of phase 2 (by month 6), the initial UI/UX design work concludes.

- Deliverable: By the end of month 6, initial design mock-ups, wireframes and technical specifications are published, after having received 2 rounds of feedback from both end users and the development team.

Milestone: At the end of phase 3 (by month 22), development concludes.

- Deliverable: After several iterative prototypes have been deployed and tested, a new stable version of OpenRefine is released, which includes advanced/improved features for visualization, customization, reproducibility and automation of workflows.
- Deliverable: OpenRefine's Continuous Integration test suite includes tests of these features.
- Deliverable: OpenRefine's various OS installer packages include the above new version of the software.

Milestone: At the end of phase 4 (by month 24): OpenRefine's technical and end user documentation are updated.
- Deliverable: Publish dedicated written documentation on OpenRefine's documentation portal explaining the new reproducibility and automation features
- Deliverable: Publish a set of short re-usable video demos and tutorials for these features
- Deliverable: Publish updated technical documentation targeted at onboarding new technical contributors

Overarching community-focused deliverables:
- Deliverable: During this project, OpenRefine's upcoming reproducibility and automation features are presented during at least 5 public activities targeted at the biomedical research community, of which at least 2 are external to the project (such as external conferences and masterclasses).
- Deliverable: By the end of this project, a representative from the biomedical research community has joined either OpenRefine's steering or advisory committee.

8. Existing Support:

List active and recently completed (previous two calendar years) financial or in-kind support for the software project(s), including duration, total costs in USD, and source of funding. Include any previous funding for these software projects received from CZI outside of the EOSS program (maximum of 250 words).

- Internships funded by Google Summer of Code (2020)
- Wikimedia Foundation grant to extend OpenRefine with functionalities for Wikimedia Commons - USD 150,000 (2021, 16 months)
- CZI EOSS Diversity grant - USD 120,000 per year (2021, 24 months)
- Google donation - USD 10,000 (2021, one time)
- NFDI grant for further Wikibase support in OpenRefine, EUR 10,000 (2022, 6 months)
- Antonin Delpeuch is employed as postdoctoral researcher at the Institut für Angewandte Informatik at the University of Leipzig (at 0.6 FTE). This EOSS5 grant will add 0.4 FTE (as an OpenRefine contractor), enabling him to work full time on this project.

9. Landscape Analysis:

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). (auto-filled from LOI; update if needed)

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 graphical user interface, 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 visualisation
- 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 and stable projects in this domain. The project has seen wide adoption for over 10 years; its user base, community and functionalities have been steadily growing. 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.

10. Value to Biomedical Users:

Describe the expected value of the proposed work to the biomedical research community (maximum of 250 words). (auto-filled from LOI; update if neeeded)

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](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

11. Category:

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

Category 1: Data management and workflows
Category 2: Machine learning and data analysis

12. Diversity, Equity, and Inclusion Statement:

Advancing DEI is a core value for CZI, and we are requesting information on your efforts in this area. Describe any efforts the software project(s) named in this proposal have undertaken to increase diversity, equity, and inclusion with respect to their contributors and audience. Please see examples from applications funded in previous cycles (maximum of 250 words)

OpenRefine has a broad user community, remarkably diverse in application domains and geographical, gender, and racial backgrounds. We are conscious that our community's diversity of background and views help us develop better technological solutions and support our user base. We prioritize diverse voices in OpenRefine's governance and community and aim to continuously evaluate our governance and community strategy (in collaboration with experts) to improve OpenRefine's accessibility and inclusive practices. OpenRefine has participated in several Outreachy and Google Summer of Code internship programs, mentoring interns and new code contributors from China, India, Cameroon and Kenya. When hiring new contractors and staff, we consciously reach out to communities that are traditionally underrepresented in open source development; for the current project on integration of Wikimedia Commons features, we work with developers from Cameroon and Costa Rica.

To support diversity in OpenRefine, we formalized and published a Code of Conduct in 2020: https://github.com/OpenRefine/OpenRefine/blob/master/CODE_OF_CONDUCT.md

Early 2022, OpenRefine hired a project director (financed via CZI's EOSS Diversity and Inclusion grant) who is tasked to increase diversity in our community and governance bodies. Via the same EOSS DEI grant, we are improving our code base and clustering algorithms to increase OpenRefine's usefulness for datasets in non-Western languages and to generally eliminate cultural biases from the tool.

OpenRefine is fiscally sponsored by Code for Science and Society (CS&S). CS&S is an equal opportunity employer committed to hiring a diverse workforce at all levels of the organization.

Optional Attachments (optional)

Attachments should be uploaded in a combined single PDF. This section can include figures, charts and tables, references for the proposal, or any additional material in support of the proposal (maximum of three pages). Uploading any additional information is optional.

Budget Description

Upload in PDF format; budgets can be uploaded in a combined single PDF or one PDF for each software project (one page per software project maximum)

Description of the costs to be funded by this grant at a high level and in narrative or tabular form, outlining costs for personnel (including names, if known), supplies, equipment, travel, meetings/hackathons/sprints, subcontracts, other costs, and up to 15% indirect costs (excluding equipment and subcontracts).
Indirect costs are limited to up to 15% of direct costs and are included within the annual budget total. Indirect costs may not be assessed on capital equipment or subcontracts, but subcontractors may include up to 15% indirect costs of their direct costs. Non-charitable entities must include a clear allocation and explanation for any indirect costs included in a proposed budget.
Budget should be requested in US dollars.
International grantees must use all grant funds exclusively for activities conducted outside the United States of America. Travel expenses to the United States (including round-trip tickets) should not be covered from the requested grant funds.
Application budgets must reflect the actual needs of the proposal. The Chan Zuckerberg Initiative will work closely with successful applicants to arrive at a mutually acceptable budget after review.

OpenRefine: Improving Reproducibility / CZI EOSS5 grant proposal		
Nov. 2022 - Oct. 2024 - Budget per year		
Expense (per year)	**Cost**	**Notes**
Development	$63,000.00	Note 1 / Note 2
Design & design research	$9,600.00	
Training and documentation	$10,000.00	Note 3
Product Management	$25,920.00	Note 4
Community outreach	$10,000.00	Note 5
Travel, meetings, equipment and PR	$12,000.00	
Code for Science & Society fiscal sponsorship fee (15%)	$23,250.00	Note 6
Yearly total	$153,770.00	
Note 1: Approximately 75% of this amount covers 0.4 FTE (as an OpenRefine contractor) for Antonin Delpeuch, OpenRefine's technical lead. In addition, 0.6 FTE will be carried out by him in his capacity of postdoctoral researcher at the Institut für Angewandte Informatik at the University of Leipzig (external funding, already covered), allowing him to work full time on this project.		

Note 2: The remaining ~25% of this amount covers various contracting tasks, including the creation of new components to OpenRefine's technical test suite, and ongoing updates to OpenRefine's packaging for various operating systems.
Note 3: This amount covers both the production of dedicated training materials (including tutorial videos), and the hours for a dedicated technical writer.
Note 4: This budget line covers 0.3 FTE for Sandra Fauconnier, OpenRefine's project director, who will act as product manager for this project.
Note 5: For community outreach, we intend to engage a contractor with good knowledge of the use of scientific datasets in the biomedical domain, acting as a community liaison and ambassador for this project.
Note 6: This budget line covers the contribution by Code for Science & Society, OpenRefine's fiscal sponsor, providing financial administration, accounting, administrative and strategic support.

Open Source Software Project(s) Details

Provide details and metrics for each open source software project that will be supported by the grant to help us assess its impact and quality.
How many software projects involved in your proposal (up to five)? If multiple software projects are involved, details must be entered for all of them where indicated below.

SOFTWARE PROJECT #1: Details:

Complete the following table for Open Source Software Project #1 of your proposal. All URLs should be in the format https://example.com and only one primary link should be provided.

Software Project name (required) OpenRefine
Main code repository (e.g. GitHub URL) (required) https://github.com/OpenRefine
Homepage URL (required) https://openrefine.org
Social media handles (if applicable) https://twitter.com/openrefine
Do you or software project key personnel have commit rights to the code repositories for this software project? (required) Yes
Short description of software project (200 words maximum) (required)

OpenRefine is a power tool to clean up messy data. Requiring no knowledge of a programming or query language, it lets users find and fix inconsistencies interactively, match their data to external databases, pull additional data from these, and many other useful operations. The resulting workflows can be extracted and applied on other projects, offering a primitive form of reproducibility that this grant proposal aims to improve on.

OpenRefine was originally designed as an Export-Transform-Load tool to populate Freebase, under the name "Freebase Gridworks". It was then briefly a Google product which became an

open source project when Freebase was discontinued. Since then, a team of users and developers has been taking care of the project, with the support from various funders.

Our very active user community runs training workshops about the tool in many universities, newsrooms and libraries. It is easy to get quick and high quality support on our mailing list, StackExchange or Twitter thanks to our committed team of trainers.

SOFTWARE PROJECT #1: List of Key Personnel:

Key personnel are people involved in the software project who will be supported by the grant if the application is successful.

Complete the following for the key personnel on the open source software project #1 (up to 5) (required); enter n/a if any field is not applicable. Personnel to be hired that have not been identified at this time can be listed in the budget section. You may need to use the scroll bar at the bottom of the table to scroll right to view and to complete all fields. Alternatively, you can tab to move through and complete the fields. To add another person/row (up to five), click the box at the end of the row.

First name	Last name	Email address	Current employer/Affiliation	Job title	Developer username if applicable (e.g., GitHub handle)	Country of Residence
Sandra	Fauconnier	sandra@openrefine.org	Code for Science & Society	Project director	trnstlntk	Netherlands
Antonin	Delpeuch	antonin@delpeuch.eu	Institut für Angewandte Informatik	Technical lead	wetneb	Germany

SOFTWARE PROJECT #1: Metrics- Quality (required):

Complete for the open source software project #1.

1. SOFTWARE PROJECT #1 : What is the software project license?

- **Permissive license** (e.g. **BSD 3-Clause**, MIT, Apache 2.0)
- Copyleft license (e.g. GPL, LGPL)
- Another OSI-approved license
- Custom license / other (please specify)

2. SOFTWARE PROJECT #1 : What is the main programming language?

- C/C++
- **Java**
- Javascript
- Python
- R
- Other (please specify)

3. SOFTWARE PROJECT #1: Does the software project have a code of conduct?

Yes. Link (optional):
https://github.com/OpenRefine/OpenRefine/blob/master/CODE_OF_CONDUCT.md
No

4. SOFTWARE PROJECT #1: Does the software project have end-user documentation?

Yes. Link (optional): https://docs.openrefine.org/
No

5. SOFTWARE PROJECT #1: Does the software project have an issue tracker?

Yes. Link (optional): https://github.com/OpenRefine/OpenRefine/issues
No

6. SOFTWARE PROJECT #1: Does the software project have a community engagement / Q&A forum (self-hosted, on Stack Exchange etc.)?

Yes. Link (optional): user mailing list at https://groups.google.com/g/openrefine
No

7. SOFTWARE PROJECT #1: Does the software project have contribution / coding guidelines?

Yes. Link (optional):
https://github.com/OpenRefine/OpenRefine/blob/master/CONTRIBUTING.md /
https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Developers
No

8. SOFTWARE PROJECT #1: Is there a corresponding package available in a package manager (PyPi, CRAN, etc.)?

<mark>Yes</mark>. Link (optional): https://packages.debian.org/bookworm/openrefine, https://formulae.brew.sh/cask/openrefine, https://snapcraft.io/install/openrefine/ubuntu
No

9. SOFTWARE PROJECT #1: Does the software project support continuous integration for testing?

<mark>Yes</mark>. Comment (optional): Java unit tests (TestNG) and Cypress UI tests
NoSOFTWARE PROJECT #1: Metrics- Impact (optional):

Complete the following for the open source software project #1. Providing metrics is optional and metrics can be approximate. For each metric, please provide a source, clarify how the metric was computed, and/or provide any other comments. For monthly metrics, please provide data from the most recent month for which the corresponding metric is available.

1. SOFTWARE PROJECT #1: Complete the following table. List the number and explanation for each, if needed:

	Number	Comment
Scholarly paper(s) (including preprints) citing or mentioning the software project	5020	According to Google Scholar as of 2022-05-25, searching for OpenRefine OR "Google Refine" OR "Freebase Gridworks"
Monthly users, if applicable (based on one or more of the following: monthly downloads from websites, monthly downloads from package managers, monthly unique requests for updates, etc.)	48,290	Total number of downloads of OpenRefine's current stable release (v3.5.2) since its release on 2022-01-26
Software projects that depend on the project	N/A	As a web-based tool designed for end-users, OpenRefine is not meant to be used as a software dependency yet. This is likely to change after this project.
Monthly visitors to project's website, discussion forum (e.g. Stack Overflow)	3,279	3,279 visits to OpenRefine's GitHub repository between May 19 and Jun 1, 2022 (statistics provided through

		GitHub Insights, which gives info about the last 14 days). We do not track visits to our website or mailing list.

2. SOFTWARE PROJECT #1: List of software projects to which key personnel on this proposal are contributing.

- OpenRefine itself
- Gephi (https://github.com/gephi/gephi)
- Wikidata-Toolkit (https://github.com/Wikidata/Wikidata-Toolkit)
- Reconciliation test bench (https://github.com/reconciliation-api/testbench)
- Reconciliation service for Wikibase (https://github.com/wetneb/openrefine-wikibase)
- EditGroups (https://github.com/Wikidata/EditGroups)