# 1. Proposal Title : OpenRefine Community and Development Support

*(maximum of 75 characters, including spaces)*

# 2. Funded Proposal ID: 2019-207403, EOSS-0000000332

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

yes

# 4. Proposal Purpose :

*One sentence (maximum of 255 characters including spaces)*
To improve how OpenRefine supports and empowers our contributor community, continue to build partnerships, and continue to make fundamental improvements to OpenRefine's architecture.

# 5. Amount Requested :

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

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

# 6. Proposal Summary/Scope of Work :

*A short summary of the application (maximum of 500 words)*

The 2019 EOSS grants allowed the project to refactor its back end to support larger data and to significantly grow its community over the last two years. Today, we have reached the core team's capacity to engage with the community and empower new contributors consistently. This new proposal will help us dedicate resources for community development while making fundamental improvements to OpenRefine architecture.

**1 Support the OpenRefine community and ensure the project's sustainability.**

We know consistency and timely response in open source is critical to attract and retain contributors. In 2020 we doubled the number of code contributors on GitHub. We have reached the limits of our core team's capacity to ensure that every question and request is addressed, that potential contributors are directed to the right resources, and that pathways to long-term contribution are maintained and grown.

We also have identified key communities and institutions to be cultivated as long-term partners for OpenRefine. It is an on-going effort to maintain discussions and engage in technical planning with partners and stakeholders to bring partnerships into being. To enable these partnerships' technical side, we will fund dedicated time from a senior technical contributor to provide timely technical support and technical opinions on new feature requests, review and merge pull requests, and prepare new releases. Bringing on a Community Engagement Lead and a senior technical role will strengthen our team's ability to develop these important relationships that are favourable to our long-term sustainability, continue to expand our ability to mentor emerging talent through internships, calls for community proposals, and stipends.

**2 Continue to improve OpenRefine’s architecture**

The grant from the first cycle of EOSS allowed us to make substantial improvements to OpenRefine architecture. Over 2020 we recognized the importance of the following further work:

2.1 Refining collaboratively
OpenRefine is designed to be run locally by the user. Although it can be hosted on a server, it is not designed for collaborative work. As operations are applied in sequence to the project, working simultaneously on disjoint parts of a dataset is rarely viable. The tool currently does not even have a notion of "user," which would let it track who performed each change.

2.2 Analyzing, sharing and reusing workflows
The ability to extract workflows as JSON objects and reapply them on other projects is a flagship feature of the tool. However, it has serious limitations. It is hard to understand what a workflow does by looking at its representation in JSON or the project history in the tool itself. There is no simple way to reorganize a workflow, isolate reusable parts or undo selected operations buried in the history.

2.3 Running workflows in production
Once a workflow has been created, one could want to run it periodically as part of a wider pipeline. Although many of OpenRefine's operations can be easily parallelized, there is no simple way to run them on data streams discovered progressively. The scheduling of operations is also naive, as they are executed in sequence without any time-sharing.

# 7. 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 projects 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 organized into three categories:

1. Spreadsheets offer an entry-level interface to the data but are time-consuming and do not scale.
2. Programming languages like Python and R offer flexibility but have a steep learning curve for non-technical people.
3. Data preparation software like OpenRefine addresses the growing data literacy gap by lowering the technical skills needed to normalize and prepare data. OpenRefine empowers those who understand the context in which the data are generated or used. OpenRefine is one of the most mature (in terms of community and functionality) open-source projects in its domain.

Other proprietary solutions include:

- Trifacta https://www.trifacta.com/start-wrangling/
- 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/

Other open source 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

A snapshot of OpenRefine through statistics from Github

- 1.5k forks
- 8k stars
- 98 contributors on Github - 46 actives in the last 12 months
- 679 pull requests submitted, 531 merged in 2020
- In 2020, the community opened 483 issues and closed 314

Our Community

- At least 146 training events provided in 2020
- At least 180 videos and tutorials published by the community in 2020
- 1,687 members on the general mailing list
- 184 members on the developer mailing list

## 8. Value to Biomedical Users :

*Briefly described the expected value the proposed scope of work will deliver 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. As a leading open source power tool to work with messy data, it is taught in countless courses and workshops around the world. 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 biomedical research alone, OpenRefine is cited in hundreds of scientific articles in genomics, Alzheimer's disease, infectious diseases, oncology, and clinical data management. In 2020 and 2021 alone, OpenRefine was cited by the following publications:

- 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

To keep OpenRefine thriving in the coming years, we want to improve how we engage and develop our contributor community, how we build lasting partnerships, and undertake fundamental improvements to its architecture.

## 9. 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. 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. If multiple software projects are involved, details must be entered for all of them. All fields are required. All URLs should be in the format https://example.com and only one primary link should be provided where requested :*

- a. Software project name
- b. Homepage URL
- c. Hosting platform (GitHub, GitLab, Bitbucket, Other)
- d. Main code repository (e.g. GitHub URL)
- e. Short description of the software project (maximum of 100 words)