1. Proposal Title: OpenRefine Community and Development Support

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

Yes, EOSS round 1. Funded proposal id: 2019-207403, EOSS-0000000332.
We are also applying to the Diversity and Inclusion grant programme.

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

This is a continuation of our EOSS round 1 grant, 2019-207403, EOSS-0000000332.

3a. Progress Report:

Provide a short summary of progress towards the deliverables in your currently funded proposal (maximum of 250 words)

The 2019 EOSS grants allowed the project to make significant progress toward sustainability at multiple levels:

By joining CS&S as fiscal sponsor, we strengthen our governance by creating an advisory and steering committee. This transition was also the opportunity to update our governance structure.

We rolled out a new user manual which replaced our GitHub wiki. This was very well received by the community and it attracted new contributors who provided edits to this documentation. Those contributions are now recorded in GitHub, on par with code contributions.

We significantly improved our contributor experience and were able to double the number of contributors by
- providing a contributor guide
- having a full-time developer on the project to answer questions and review contributions promptly.
- participating in the Google Summer of Code and Outreachy internship programs.
- improving our test suite to catch bugs earlier, thanks to a comprehensive integration test suite using the Cypress framework.

Finally, we refactored our back end to support larger datasets. OpenRefine workflows can now be executed on three different executors (local, Spark and testing), depending on the needs.

This new architecture will be released in the next major release series of OpenRefine, 4.x. The scalability has already been successfully tested on use cases reported by our users, where the original backend failed.

4. Proposal Purpose (required):

Autofill from LOI

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 (required):

- Year 1
- Year 2
- Total: USD 400,000

6. Proposal Summary (required):

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

The 2019 EOSS grants allowed the project to refactor our back end to support larger datasets and significantly grow our user and contributor communities. This growth is critical to solidifying OpenRefine's pipeline of new contributors and long-term project sustainability. To maintain consistent community engagement and develop new contributors across our large community, we will dedicate resources to community development. This new proposal will help us dedicate resources for community development while building on fundamental improvements to OpenRefine's 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. With this growth, brought on by both new users and our community engagement strategy implemented in the EOSS 1 grant, we have reached the limits of our core team's capacity. It is critical for OpenRefine 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.

In this proposal, we will adjust how OpenRefine approaches staffing by introducing a project director role separate from the technical lead. This change will help us better match the various sustainability, partnership, contributor, and community needs of the project with our team's capacity and expertise.

Through the course of our EOSS 1 grant, we have identified key communities and institutions to be cultivated as long-term partners for OpenRefine. It is an ongoing 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. Rather than relying on technical staff to develop and manage software and project strategy/partnerships, OpenRefine will resource a project director and a separate senior technical role. EOSS 1 allowed us to lay the groundwork for a sustainable OpenRefine, including bringing on new funders and partners. Looking ahead, strong strategic and technical leadership will strengthen our team's ability to develop important relationships favorable 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's architecture, including the ability to work with large datasets by revamping our data model and integrating with Apache Spark. Over 2020 we recognized the importance of the following further work to address long-standing items on our roadmap, namely providing our user the ability to collaborate on a project and making our workflow engine more robust. Those improvements are based on Antonin Delpeuch's publication *A complete language for faceted dataflow programs*. In arXiv:1906.05937 [Cs, Math], in Proceedings of the Applied Category Theory 2019, Oxford, June 2019.

7. Work Plan (required):

A description of the proposed work the applicants are requesting funding for, 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)

With this grant, we will focus on the following areas.

Hiring a project director. We want to find someone who embodies the diversity of OpenRefine's user community and entrusts them with developing and executing the project's strategy and governance. OpenRefine is an open-source project with a small core team, where developers have a disproportionate influence on the project's priorities. Hiring a project director with a strong connection to the user community can help balance out this influence and be a driving force to onboard project members from broader backgrounds. This role will be supported by the project's existing advisory and steering committees, and will be charged with running the project on a day-to-day basis, building out OpenRefine's governance, and leading its strategy in the long term:

- Represent the project publicly, liaising with partners
- Updating the project's roadmap, hand in hand with the steering and advisory committees
- Fundraising and reporting to funders
- Organize the project's participation in internship programs
- Update governance, code of conduct, and contributing documents and create a safe and welcoming space for contributors.

Hiring a tech lead. Over the past few years, we have worked on grants which focused on specific improvements to the tool. This has drawn core team members away from basic day-to-day maintenance tasks that are less visible but crucial to the project's health. For instance, triaging bug reports, reviewing pull requests, and releasing new versions are time-consuming processes that were not directly covered by our funded developer roles. Therefore we have identified the need to allocate resources to those tasks as well. This maintainer role will be a part-time responsibility, which will be scaled dynamically to accommodate varying activities (as it requires more effort in the run-up to internship programs, for instance). This will improve the experience of contributors, for whom swift reviews are critical to staying motivated, and of users who request more frequent releases. Establishing a technical lead role with a high-level view of the project's technical position will also support the project director in planning for new partnerships and grants.

Continuing architectural improvements. The EOSS 1 grant allowed us to introduce fundamental changes to OpenRefine's representation of project data, which makes it possible to work on datasets which do not fit in memory. We now need to tackle the follow-up improvements that this change makes possible. The OpenRefine Technical Lead will be responsible for organizing the delivery of the following items with the support of the community and paid developers.

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. Team and collaborative science are increasingly common approaches to today's complex biomedical research challenges. Allowing users to "refine collaboratively" will resolve a common feature request and enable teams to take full advantage of OpenRefine. Providing a native hosted version of OpenRefine will also address long-standing issues for users within organizations limiting what program can be installed on their computer.

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. Easy sharing and reuse of workflows will help OpenRefine to be a part of reproducible research practices, which are becoming the norm in science. By improving the experience of analyzing, sharing, and reusing OpenRefine workflows, we can improve our user experience and support reproducible research best practices.

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. Improvements to how workflows are run in production will support scientists' diverse workflow needs.

8. Milestones and Deliverables (required):

List expected milestones and deliverables, and their expected timeline. Be specific and include (where possible) any goals for metrics the software project(s) are expected to reach upon completion of the grant (maximum of 500 words)

Milestone: Hire and onboard a project director from the OpenRefine user community by October 2021. Deliverables for this grant focus on the roadmap, governance, and partnerships. This role would also be funded by the EOSS D&I grant (if successful), and deliverables for that portion of the project director's role focus on diversifying the new contributor pipeline.

1. **Deliverable:** within six months after the hire, the project director releases a 2-year roadmap with a plan for revision and updates.

2. **Deliverable: Increased stakeholder representation in governance.** The project director will develop OpenRefine's existing governing bodies over the next 24 months, focusing on bringing users and institutions into the project's governance. This deliverable links to the EOSS D&I proposed deliverable of increasing diversity along geographic, racial, and ethnic axes in governance.

3. **Deliverable: Increased revenue diversity.** The project director, supported by OpenRefine's governing bodies, will set a goal of agreements with 2-4 new institutional partners and 2-4 new funders (fee-for-service, grant, donation, etc.) over the next 24 months.

Milestone: Hire a technical lead director from the OpenRefine user community by December 2021. Deliverables for this grant focus on the upkeep of OpenRefine project. This role will also support our EOSS D&I grant (if successful) by providing mentorship to our interns.

1. **Release frequency:** The technical lead will define a release schedule with at least two

releases per year.

2. **Review Pull Request, Triage tickets, and answer technical questions:** the technical lead will answer in timely manner questions, issues, and proposed contributions by the community.

3. **Maintain OpenRefine technical infrastructure,** including update dependencies, patch security issues, and our continuous integration infrastructure

Continuing architectural improvements.

1. **Identifying resources by March 2022**: the Technical Lead identified and recruited new developer(s) to support architectural improvements.

2. **Workflow visualization - 5 months** We add diagrammatic representations of the list of implemented operations, letting users better understand the structure of their workflows.

3. **Operation reordering - 3 months** We introduce the possibility of reordering operations in the undo/redo history. This feature would only be available for row-wise operations working on different columns, implementing the new interface. Reordering could be triggered directly by dragging nodes on the graphical visualization.

4. **Concurrent operations - 3 months** We add the possibility of running independent operations concurrently, such as reconciling two different columns in parallel. The workflow representation should also make it possible to execute different operations on disjoint subsets of rows.

5. **Partial computations - 6 months** We return the control to the user before the full computation of long-running operations terminates. For instance, when reconciling a column, the user can immediately inspect the first few reconciliation results. They can perform other operations while reconciliation is progressing. These other operations are executed immediately, even if they depend on the reconciliation results. This relies on the streaming capabilities of the underlying execution backend.

6. **Multi-user support - 7 months** We add a notion of a "user" to keep track of the author of each change and make necessary adjustments to the interface and server to let multiple users work on the same project seamlessly.

9. Existing Support (required):

List active and recent (previous two calendar years) financial or in-kind support for the software project(s), including duration, amount in USD, and source of funding. Include in this section any previous funding for these software projects received from CZI (maximum of 250 words)

In the past two years, OpenRefine has been supported by:

- EOSS round 1 grant (2019) USD 200,000
- Internships funded by the Google Summer of Code (2020)
- Wikimedia Foundation grant (under review, 2021) USD 100,000
- EOSS Diversity grants (under review, 2021) USD 120,000 per year for two years

In addition to this OpenRefine now receives support in the form of developer time, user support, and governance participation from independent individuals and from people based at institutions including RefinePro and The Carpentries. The project director will work to formalize in-kind relationships with institutions.

10. Landscape Analysis (required):

*Briefly describe the other software tools (either proprietary or open source) that the audience for this proposal is primarily using. How do the software projects in this proposal compare to these other tools in terms of size of user base, 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 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

11. Value to Biomedical Users (required):

Briefly describe the expected value the proposed scope of work will deliver to the biomedical research community (maximum of 250 words) ***(auto-filled from LOI; update if needed)***

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.

12. Diversity, Equity, and Inclusion Statement (required):

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.

To support diversity in OpenRefine, we formalized and published in 2020 our code of conduct available here:
https://github.com/OpenRefine/OpenRefine/blob/master/CODE_OF_CONDUCT.md

As stated in the application, the project director (financed via the EOSS-4 and EOSS Diversity and Inclusion grants) will be tasked to increase diversity in our community and governance bodies.

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 thereby creating a culture that allows us to better serve our clientele, our employees and our communities. We value and encourage the contributions of our colleagues and strive to create an environment where everyone can reach their full potential and drive outstanding results. All qualified applicants will receive consideration for employment without regard to race, national origin, age, sex, religion, disability, sexual orientation, marital status, veteran status, gender identity or expression, or any other basis protected by local, state, or federal law. This policy applies with regard to all aspects of one's employment, including hiring, transfer, promotion, compensation, eligibility for benefits, and termination.

Budget Description

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.

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.

EOSS-4	
September 2021	
Item	**Amount**
CS&S Management fee (15%)	$30,000.00
Project Director	$45,000.00
Technical Lead	$90,000.00
Technical Contractors	$25,000.00
Travel (if permitted)	$5,000.00
Miscellaneous	$5,000.00
Total	$200,000.00
September 2022	
Item	**Amount**
CS&S Management fee (15%)	$30,000.00
Project Director	$45,000.00
Technical Lead	$90,000.00
Technical Contractors	$25,000.00
Travel (if permitted)	$5,000.00
Miscellaneous	$5,000.00
Total	$200,000.00

Through this grant we plan to finance three positions. We have early discussion regarding potential candidates but we have not yet identified the individual for each position.

1. Project Director Role
2. Technical Lead
3. Technical Contractors

We are currently applying to three different grants (grants are in review from EOSS Diversity and Inclusion, the Wikimedia Foundation, and in development to the Institute of Museum and Library Services), the outcome of those application will impact how we finance the following two positions:

Project director We expect a budget of USD 70k-100k per year for a full-time role, and plan to cover these costs with funding from multiple sources. We identified Sandra Fauconnier as a potential candidate. Sandra is the project coordinator for the Wikimedia grant. If we are unable to secure funds to cover this salary, we will either reduce the scope for a project director for a part-time role or merge with the technical lead (depending on the profile of our lead developer).

Technical Lead: We expect a budget of USD 90k for this position solely financed by the EOSS-4 grant. We would like to offer the position to Antonin Delpeuch in continuation of his contract since 2018 (financed via the Google New Initiative donation and EOSS-1 grant).

Technical Contractor: We allocated a budget of USD 25k to USD 30k per year from the EOSS-4 grants to hire developers and help with the implementation of the architectural improvement. We identified the following scenario depending on other grants outcomes and profile of technical lead:

1. Hire a part time developer for the course of two years
2. Extend the contract of the OpenRefine Developer from the wikimedia grants. In that case we will have a full time developer for a year starting in Summer or Fall 2022.