# 1. Proposal Title : OpenRefine for Everyone

*(maximum of 75 characters, including spaces)*

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

# 3. Proposal Purpose:

*One sentence (maximum of 255 characters including spaces)*

Continuing to diversify our contributors by building capacity in project management and eliminating cultural or linguistic biases in the tool itself.

# 4. 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 240k total, split 120k USD per year for two years.

# 5. Proposal Summary/Scope of Work:

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

OpenRefine has a broad user community, which is already remarkably diverse both in terms of application domains and of geographical, gender, and racial backgrounds. A strong network of enthusiasts trains newcomers and a committed team of translators maintains the tool in many languages. However, like many software projects, we lack such a diversity among code contributors.

Through 2020, we went through tremendous growth as we doubled the number of active contributors, increased the number of languages translated, and continue to see more users relying on OpenRefine for their research. We participated for the first time in the Outreachy and the Google Summer of Code programs, both of which attracted talented code contributors from underrepresented backgrounds. Outreachy is a diversity initiative that provides paid, remote internships to people subject to systemic bias and impacted by underrepresentation in the technical industry where they are living. While this was extremely beneficial to the project, we could unfortunately not participate again in summer 2021 as we lacked mentors to support our interns. This made clear that we need to build capacity to support these initiatives, which requires dedicated resources. To participate in these initiatives we need staff to not only mentor selected interns, but also to coordinate project's participation in these programmes (coordinating what work will be done with mentors and interns, managing postings in intern programs, supporting interns to return as contributors), and sustain a welcoming atmosphere for

prospective and existing contributors.

We have also identified a range of areas where the tool is still biased towards certain communities. For instance, some parts of the interface still display English text that cannot be translated through our crowdsourcing platform, due to architectural issues. Beyond the interface, this bias also affects key features of the tool, which have been built with a western mindset. For instance, the reconciliation algorithm relies on some heuristics designed specifically for English. Similarly, many clustering features are centered around the Latin alphabet, making them of little use for many datasets. Even basic tools such as number or date parsing utilities only support a narrow range of formats, with similar biases. More broadly, the conversation around the latent assumptions of data science has gained some momentum (for instance with Data Feminism by D'Ignazio and Klein, or Ruha Benjamin's work) and the points it raises are of course relevant for OpenRefine. We want to tackle these issues and we think this will be a great opportunity to not only serve our users better but also attract contributors from diverse backgrounds, who will be in a good position to identify other such biases in OpenRefine.

# 6. Work plan

*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. Please specify how the proposed activities will be structured and organized in the context of advancing the participation, retention, and leadership progression of contributors that are systemically underrepresented in scientific open source. Provide information on how this work will fit with the open source project's roadmaps and ongoing initiatives, and their target audience. Please describe any previous record of related efforts by the applicant/key personnel of the proposal. If this proposal involves participation from another organization, and/or multiple open source projects, please describe the collaboration and respective responsibilities. (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 entrust them with the reins of the project. OpenRefine is an open source project with a small core team, where developers have a disproportionate influence on the project's priorities. We think 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.

Within the first six months, the project director will develop an assessment plan to measure the impact of this grant on our community, in collaboration with CS&S and outside experts. This will help us learn about our community and measure the impact of the activities (internships, community management, etc) on our community members and the project as a whole.

The Project Director will share most responsibilities of the advisory committee, running the project on a day-to-day basis and leading its strategy on 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 programmes
- Update governance, code of conduct and contributing documents and create a safe and welcoming space for contributors.

We count on a budget of USD 70k-100k per year for a full-time role, which would be jointly funded by other grants (EOSS 4, potential funding from the Wikimedia Foundation and the Institute of Museum and Library Services). If we are unable to secure complementary funds to cover this salary from other sources, we will direct the Diversity Grant to finance a leadership role to mentor students and address cultural biases in the tool as described below).

**Securing the project's participation in internship programs**. Our experience with these programs is that they require dedicated resources, for two reasons. First, they incur a big workload on the project team ahead of the internships themselves, with a flock of prospective contributors coming to the project and needing specific help in their first contributions. Second, mentors need to be available to support the interns throughout their projects. Therefore, we want to set aside a budget to compensate mentors, in addition to the intern stipends (that the Outreachy program requires the projects to fund).
Potential mentors include Tom Morris (with long term experience of OpenRefine as a contributor and of Google Summer of Code as a mentor), Antonin Delpeuch (who coordinated OpenRefine's participation in GSoC and Outreachy in 2020), Lu Liu and Lisa Chandra (past Google Summer of Code interns), Ekta Mishra (past Outreachy intern).

**Tackling cultural biases in the tool.** With fewer cultural biases, we expect to see more users and contributors with diverse backgrounds coming to the project. Our Project Director and Technical Lead (hired thanks to the EOSS-4 grants) will ensure a smooth onboarding and develop relationship with those new users. This work will be proposed as internship subjects, as this front consists in many relatively small and independent tasks, such as:

- Improvements to date parsing with better support for non-Western date formats
- Improvements to number parsing with locale support
- Generalization of the reconciliation API to expose service-defined features
- Localization of parts of the UI which cannot be translated yet for architectural reasons
- Implementation of clustering heuristics suitable for non-latin alphabets
- Better support for legacy encodings

We will leverage our community of translators to identify more aspects of the tool which could be better adapted to other cultures. For instance, we have had fruitful interactions with the OpenRefine community in Japan which identified language-specific issues. This was then the motivation for making clustering features extensible, which was released in OpenRefine 3.2.

# 6. Milestones and deliverables

*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)*

By the end of the grant we want to:

1. **Milestone: Hire and onboard a project director from the OpenRefine user community by October 2021.** Deliverables for this grant focus on diversifying the new contributor pipeline. This role would also be funded by the EOSS 4 (if successful) and deliverables for that portion of the project directors role focus on the roadmap, governance, and partnerships.

    a. **Deliverable: Regular participation in programs that develop OpenRefine's contributor pipeline.** Project director and technical lead (hired thanks to the EOSS-4) will work together to manage OpenRefine's participation in at least three rounds of internship programs (GSOC, Outreachy, or others) with a goal of cultivating experienced interns who return as volunteer contributors or for contract work. This will support junior developers from diverse backgrounds by mentoring them through the process of making a significant contribution to an open source project. We want to encourage as many of these interns to keep contributing to OpenRefine afterwards but that is not in itself our primary goal, since our alumni can be active in other projects, so their training is a service to the community. **Timeline**: One to two rounds of internships per year for the duration of the grant, each lasting for a few months.

    b. **Deliverable: Increased diversity along geographic, racial, and ethnic axes in governance.** The new project director will develop OpenRefine's existing governing bodies over the next 24 months, yielding increased diversity in multiple axes on OpenRefine governing bodies. Part of this deliverable include updating and maintaining our governance, code of conduct and contributing documents to create a safe and welcoming space for contributors. This deliverable links to the EOSS 4 proposed deliverable of increasing stakeholder representation in governance. The project director will be supported by the existing committees to increase diversity in multiple axes on OpenRefine governing bodies.

2. **Milestone: Have tackled at least three cultural biases in the tool**. These will be tackled either during the internships or by independent efforts. Timeline: flexible, as we want to leave our interns a choice in the projects they tackle.
    a. By December 2021: Assessment of cultural biases in OpenRefine with the community
    b. By September 2023 correction of at least three biases identified previously.

# 7. Expected Outcomes:

*Please describe the impact and expected outcomes of the proposed activities on the* participation, retention, and leadership *progression of contributors from underrepresented groups in your project(s). Top-tier applications will include clear and bold hypotheses about outcomes resulting from the work. (maximum of 250 words)*

With this grant, we aim to achieve the following outcomes:

- Develop an assessment plan to measure the impact of this grant on our community.
- Attract new contributors and further develop the pathway to leadership in OpenRefine by:
    - Securing OpenRefine's participation in internship programs (Outreachy, Google Summer of Code), which will attract at least 10 new contributors from underrepresented backgrounds to the project.
    - Building up the mentoring and administrative capacity. Onboard former interns and recent contributors in leadership roles (advisory committee, steering committee, maintainer roles). We want to entrust at least 5 new contributors with maintainer rights.
    - Develop and document governance processes to support onboarding and leadership development.
- Develop a stronger sense of community among contributors encouraging long-term commitment to the project by:
    - Running regular calls or events where contributors can share skills, present use cases and provide feedback regarding their usage.
    - Provide training to project leaders around inclusion and code of conduct enforcement.
- Initiating exploration into the tool's cultural biases by:
    - Educate our community on bias in open source processes, software, and user experience
    - Engage our user and contributor community to identify and develop strategies to address bias in OpenRefine

# 8. Evaluation and learning goals

*Please describe how you will assess and monitor progress towards the desired outcomes and what you expect to learn about the funded activities by the end of the grant period. Consider including any qualitative and quantitative methods you plan to use to asse*ss progress towards the desired outcomes, and considerations on future scalability a*nd requirements to further sustain the work funded by the grant, if activities are expected to continue past the duration of the grant. Please refer to the expected impact and learning goals section in the LOI announcement. (maximum of 500 words)*

We want the project director to be genuinely able to emancipate from the current advisory committee which hires them in the first place. This is necessary for the project director to be in a

real leadership position. This means proactively countering the initial subordination relation which might be strengthened by the fact that we want to hire someone from a less technical background than the current advisory committee. With the support of Code for Science and Society, we will review this situation 3, 6, 12 and 18 months after the hire, to assess

- which decision-making responsibilities they have been able to take over and what prevents them from becoming more independent.
- Progress made against the Diversity and Inclusion assessment plan defined at the beginning of the contract.

Internships will be reviewed following the processes laid out by the internship programmes themselves, which provide sufficient oversight on the progress of the intern and the working relationship with their mentor. In addition to that, we want to review our own processes around the internship programmes, by having a conversation between interns, mentors and administrators at the end of each round. In 2020, such a conversation helped us identify that it is better to have a single mentor per intern, to avoid dilution of responsibilities.

For the cultural biases in the tool, we will conduct a survey advertised to our user community to let them surface the cultural barriers in the tool that we should tackle first. For instance, the dialogue with our Japanese community made it clear that biases in the application logic (clustering, number parsing) are more problematic than lack of translation support in some parts of the user interface. We believe an online survey could let us surface similar feedback from other backgrounds.

# 9. Existing support

*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-4 grants (under review, 2021) USD 400,000 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.

# 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.  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)

## 8. Budget

(Coordinated with EOSS 4)

# General feedback for LOIs

*In addition, we have general feedback based on reviewing the LOIs. Please ensure that your full proposal includes the following:*

- ***Address inclusion:*** *Some proposals are primarily focused on measures to expand the diversity of an open source project's target audience. Please make sure your proposal addresses any relevant aspect of inclusion as it relates to a project's contributors or maintainers.*
- ***Connect to open source contributions:*** *The RFA aims to support diversity and inclusion in leadership progression in open source specifically, not in STEM or academia more generally. Please make sure your proposal emphasizes open source, as our expert reviewers will be evaluating this in scoring proposals.*
- ***Be explicit about expected impact:*** *To facilitate the review of the full proposal, please state explicitly what groups/dimensions of representation are the focus of your proposal*