OpenRefine Catalyst Grant Application

1. **ABSTRACT**

OpenRefine is an open-source software that enables users with and without programming skills to perform advanced data cleaning, transformation, and reconciliation tasks that go beyond the capabilities of traditional spreadsheet tools. One of the key elements of OpenRefine is called *faceting*, which allows users to explore the content of a column in a dataset to look for patterns and trends. Facets are essentially aspects or angles of data variance in a given column. While it has been central to OpenRefine since its inception, but the time has come to reimagine faceting experience to meet modern expectation for usability, interactivity, and reproducibility.

We propose a development project centered on widely requested features like optional facet popups, datatype enhancement, recoverable facet history, and improved usability of histogram, timeline, and scatterplot facets. Together, these developments will make OpenRefine a more powerful platform for visual data exploration and open research workflows, helping a new generation of users engage more deeply with their data.

2. **TEAM**

OpenRefine is fiscally sponsored by Code for Science and Society, a 501(c)(3) charitable organization in the USA. The OpenRefine team oversees the software's development and sustainability, coordinating volunteer contributions and managing technical planning and releases. Over the past year, 22 contributors have merged more than 200 pull requests, reflecting the project's active and collaborative development model.

The core team includes Rory Sawyer, Lead Developer, and Martin Magdinier, Project Manager. Both have contributed to OpenRefine for several years and provide continuity across releases and community engagement. Rory maintains the codebase while Martin oversees coordination and partnerships.

To deliver this project, the team will collaborate with an external UX/UI designer, funded through the grant, to prototype and test the new faceting interface. This partnership will ensure that the improvements are guided by modern usability standards and community feedback.

Together, this mix of experienced maintainers, institutional advisors, and design expertise uniquely positions OpenRefine to deliver a thoughtful modernization of its faceting tool, strengthening its role as an accessible, user-focused platform for open data exploration.

3. **PROBLEM & SOLUTION**

Many OpenRefine users describe **faceting as a way of having a conversation with their data**, where each discovery leads to new questions and deeper understanding. Facet allows user to visualise and filter unique values within a field. In its current form, the faceting tool offers limited ways to group and explore text data, timelines, numbers, or find nulls and duplicates. These are useful to find outliers, detect irregularities. Yet, the faceting tool limits how far that conversation can go. Users often find themselves exporting datasets into other software for detailed analysis, breaking the flow of exploration and learning.

Several practical issues also restrict the experience: facets stack within a single side panel, which quickly becomes cluttered; numeric and date facets are rigid and lack control over binning or scale; and users cannot save or restore facet configurations between sessions. As a result, OpenRefine excels at data cleaning but falls short as an exploratory environment.

We propose to reimagine faceting as a more interactive and reproducible process through five improvements:

- **Usability and Layout:** Allow facets to open in detachable windows for multi-column exploration.
- **Visualization and Interactivity:** Redesign histograms, timelines, and scatterplots for clarity and direct manipulation of ranges and variables.
- **Datatype Awareness:** Provide lightweight statistics and highlight inconsistencies within column values depending on whether they are text, numeric or dates.
- **Reproducibility:** Enable saving and exporting facet states and history with the project.

These needs have been articulated repeatedly by users in Barcamp 2024 discussions, the 2024 feature prioritization survey, and the project's goalposts for 2024–25. They are also reflected in multiple community issues reported via GitHub issues.

By addressing these long-standing requests, this project will make OpenRefine's facets a more conversational and visual interface for data exploration, empowering researchers and GLAM professionals to gain insight directly within the open-source tool they already rely on.

4. **BUDGET**

Category	Description	Allocation
Design	Contract with a UX/UI design studio (e.g. Ura Design) for user research, wireframing, and interface prototypes of the new faceting tool.	**£5,500**
Development	Front and back end implementation. Led by the core developer over ~80–100 hours total.	**£13,250**
Project management	Coordination, reporting, and community engagement over the six-month period (approx. 7 h/month).	**£2,500**
Fiscal sponsor fee	Administrative support by *Code for Science and Society*	**£3,750**
Total		**£25,000**

Project duration: 6 months including 1-month ramp-up and activation, 1-month design, 4-month implementation and testing.

5. **PRODUCT**

OpenRefine is a free, open-source software for exploring, cleaning, transforming, and reconciling data. It provides a spreadsheet-like interface combined with powerful operations that help users detect inconsistencies, structure information, and align it with external sources such as Wikidata, VIAF, ORCID, and ROR.

Maturity: OpenRefine has been an open-source project for over 17 years, originally developed at Metaweb as *Freebase Gridworks*. It became *OpenRefine* in 2012 and continues to be actively maintained, with one major release each year and several minor updates.

Adoption: OpenRefine is a mature and widely used product, averaging about 15,500 downloads per month and cited in around 800 academic publications each year. It serves thousands of active users across research, libraries, journalism, and GLAM institutions, with an active contributor community and regular releases.

More informations.

- Website: https://openrefine.org
- GitHub: https://github.com/OpenRefine/OpenRefine
- Usage stats: https://openrefine.org/usage
- Project history: https://openrefine.org/openrefine_history

6. **PRICING**

OpenRefine is open-source software and completely free to use for everyone. The project follows a public-good model rather than a commercial one. Our value lies in accessibility, openness, and interoperability across research and cultural heritage institutions.

We sustain ongoing development through grants, institutional support, and community donations. Recent funders include the Chan Zuckerberg Initiative and the Wikimedia Foundation.

7. **LONG TERM VISION**

OpenRefine's long term vision is to help create a more informed world where working with data is easy and engaging (Mission & Vision). We believe that everyone, not just programmers, should be able to explore, clean, and share data with confidence.

This project supports that vision by improving the faceting experience so that data exploration becomes as intuitive and engaging as data transformation. By lowering the technical barrier for discovery, we help researchers, librarians, and cultural heritage professionals gain deeper insights into their data without relying on code.

In the long term, OpenRefine aims to keep the paid team small and sustainable while continuing to recruit and mentor new contributors.

8. **COMPETITORS**

Alternative tools fall into three main categories:

1. **Spreadsheet software** such as Excel or Google Sheets, which is mature and widely adopted but limited for messy or large datasets.
2. **Programming environments** such as Python and R, which are flexible and powerful but have a steep learning curve.
3. **Visualization platforms** such as Tableau or Datawrapper, which provide attractive graphics but little support for iterative data cleaning or transformation.

These tools are all established, proprietary, or code-driven products that serve specific user segments rather than bridging them.

OpenRefine is different. It combines an accessible graphical interface with the power of advanced data transformation, reconciliation, and exploration, all within a free and open-source environment. The modernization of OpenRefine's faceting tool will strengthen this advantage by bringing powerful visual exploration into the same trusted workspace, reducing the need for multiple tools and making data work more engaging for everyone.

9. **MARKET**

OpenRefine Core market segments

- Open science and research data management: Used by researchers, data stewards, and repository managers to clean, transform, and prepare datasets for sharing.
- GLAM institutions: Libraries, archives, and museums use OpenRefine to align metadata with authority files such as Wikidata, VIAF, and ORCID.
- Education and training: Included in the Library Carpentries curriculum and in other external training programs that build a strong base of advocates and trainers.
- Civic tech and data journalism: Used to clean and cross-reference public data for transparency and accountability work.

Together, these segments represent tens of thousands of users in more than 250 countries, with about 15,500 monthly downloads.

OpenRefine grows through word of mouth, institutional adoption, and advocacy from trainers and contributors who integrate it into workshops and courses.

10. **PROGRESS TO DATE**

The project operates with a low burn rate, maintaining an annual budget of approximately US $140,000 (around US $11,500 per month) to support a small part-time core team. This includes the Developer and Contributor Engagement Lead (US $7,500/month), Project Coordinator (US $2,500/month), essential software licenses, and a 15% fiscal sponsorship fee to Code for Science and Society for administration and compliance. This lean model ensures that almost all funding directly supports maintenance, community engagement, and development.

Contributor and user activity remain consistent, with ongoing releases and community participation. We expect gradual growth in both user adoption and contributor involvement, supported by the core team and by continued collaboration with trainers and institutional partners in the open-science and GLAM communities.