FOSS Sustainability Fund 2026

Project Name *

OpenRefine

What is your name? *

Martin Magdinier

What email address should we use to contact you? *

martin@openrefine.org

Requested Funding *

USD 92,500

Duration *

12 months

1. Concept Summary

Please briefly describe (in 3-5 sentences) what you hope to achieve with this funding. How will this work contribute to the sustainability, interoperability, reproducibility, security, and/or overall resilience of the internet freedom technology ecosystem? [150 word limit] *

The internet freedom technology/tool ecosystem includes, amongst other technologies, circumvention tools (like VPNs), secure messaging apps, shutdown-resilient tools, anti-censorship and privacy-preserving technologies, as well as these tools' foundational software infrastructure. If your software is infrastructure, it must demonstrate downstream dependencies in the internet freedom tool ecosystem. If it is user-facing, it must be an open source internet freedom tool and demonstrate a large user base in repressive information control contexts.

This project enables journalists and investigators to process sensitive datasets locally, without relying on external reconciliation services.

OpenRefine is used by organizations such as Bellingcat, the International Consortium of Investigative Journalists, and ProPublica to clean, match, and verify data prior to publication. These practices are representative of broader use across data journalism, Open Source Intelligence (OSINT), and civic technology communities.

By integrating reconciliation directly into the application, the project removes a key external dependency, avoiding the need to send sensitive data to third-party services and improving reliability and resilience. The work also aligns OpenRefine with the evolving reconciliation API specification, improving interoperability across data ecosystems and ensuring long-term maintainability. By embedding this functionality within the core application, the project supports reproducible, auditable data-processing pipelines for public-interest investigations, strengthening the sustainability of a widely used tool in the internet freedom technology ecosystem.

2. Eligibility Criteria Questions

2.1. Is your project a core dependency of internet freedom technologies? *

- ☐ Yes
- ☐ **No**

2.2.2 Is your project an app or other user-facing tool? *

- ☐ **Yes**
- ☐ No

2.2.2 Does it have a substantial base of active users or beneficiaries facing information controls in repressive contexts? *

- ☐ **Yes**
- ☐ No

2.3. If no to the above questions, does this proposed effort otherwise somehow impact technologies that fit the above two categories (core dependencies of internet freedom technologies, or apps/user-facing internet freedom tools that have a substantial bases of active users/beneficiaries facing information controls in repressive contexts?)

- ☐ Yes
- ☐ No

2.4 List relevant examples of internet freedoms apps and tools that your project impacts, and/or list primary users demographics/beneficiaries. [300 word limit]

OpenRefine is used by public-interest data practitioners, including journalists, OSINT researchers, civic technologists, watchdog organizations, and data literacy networks. These users often work with incomplete, inconsistent, or sensitive data. OpenRefine enables domain experts to directly clean, standardize, and reconcile datasets without relying on custom code, making these processes accessible while preserving their understanding of the data. This reduces the need for intermediary technical steps and helps ensure that data transformations remain transparent, reproducible, and aligned with the original context. In some cases, these activities take place in constrained or high-risk information environments, where the ability to process and reconcile data locally, without relying on external services, is particularly important.

In data journalism, OpenRefine is used to normalize inconsistent records, deduplicate entities, and ensure data quality before publication. For example, investigations by ProPublica and Rocky Mountain PBS (https://www.propublica.org/article/they-faced-foreclosure-not-from-their-mortgage-lender-but-from-their-hoa) used OpenRefine to standardize court records and identify patterns in foreclosure cases.

In OSINT workflows, organizations such as Bellingcat (https://www.bellingcat.com/resources/2023/08/24/following-the-money-a-beginners-guide-to-using-the-opencorporates-api) and the International Consortium of Investigative Journalists (https://offshoreleaks.icij.org/docs/reconciliation) use reconciliation workflows to match local datasets against corporate registries, sanctions lists, and leak databases.

In civic technology and watchdog contexts, organizations such as Open Ownership (https://www.openownership.org/en/blog/reconciling-beneficial-ownership-data/) and mySociety (https://www.mysociety.org/2020/07/31/screening-for-conflicts-of-interests-in-ownership-data/) use OpenRefine to reconcile fragmented public datasets, including beneficial ownership records and procurement data.

OpenRefine is also used in global data literacy initiatives, including through networks such as the Global Investigative Journalism Network and the Open Knowledge Foundation, supporting access to these workflows in regions including Latin America, Africa, and the Middle East.

2.5. How long has your project been active? *

select one

- ☐ New project (less than one year or brand new)
- ☐ 1-2 years

- ☐ 2-5 years
- ☐ 5-10 years
- ☐ **10+ years**

2.6 If your project is less than two years old, please briefly explain how this new project improves interoperability, security, maintainability, or increased long-term sustainability of actively developed, relied-upon internet freedom technologies. [200 word limit]

2.7 Briefly describe the current status of the project as relevant to long-term sustainability, such as whether it is actively developed/updated (how often are updates pushed?), dormant (needs to be revived), undermaintained, number of maintainers, etc. [200 word limit] *

OpenRefine is an actively maintained open-source project with over a decade of continuous development. The project follows a regular release cycle, targeting one major release per year, along with incremental updates. Development is conducted in the open on GitHub, with ongoing issue tracking, pull requests, and code review.

Maintenance is led by a small Core Developers Group (currently two active volunteer maintainers), supported by a broader contributor base. Over the past year, more than a dozen contributors have been active, with steady activity across issues and pull requests. This reflects a healthy but capacity-constrained project.

Reconciliation is a critical and widely used functionality, but developing and aligning it with ever-evolving standards is challenging in a volunteer-based model. Addressing the current interoperability and sustainability demands requires a dedicated and sustained effort that our volunteer development team cannot afford to make on top of their maintenance work.

Providing dedicated resources for this work will enable targeted improvements to core infrastructure while maintaining the project’s release cycle and supporting its long-term sustainability.

3. Concept Narrative

3.1 Workplan Summary: Provide a high-level overview of the proposed scope of work (objectives and activities). [500 word limit] *

Example structure:
Objective 1: Briefly describe high-level objectives with intended impact.

Activities: Briefly summarize activities needed to complete the listed objective.

Please note, this section is meant to be a high-level overview. If this Concept moves to the proposal stage, you will be asked to expand on work plan details.

This project focuses on strengthening OpenRefine's reconciliation infrastructure. Reconciliation is the process of matching messy or inconsistent data, such as names of people, companies, or organizations, to authoritative datasets in order to assign stable identifiers and enable reliable analysis. It is a core step in many public-interest data workflows, but today it relies on a mix of external services and locally deployed tools, which require additional technical setup. This creates fragmentation, increases operational complexity, and limits accessibility for users who need reliable and privacy-preserving data processing.

The work is structured into three objectives. The first two prepare OpenRefine's existing reconciliation architecture for the third: enabling local reconciliation directly within the application.

1. Improve the reliability and transparency of existing reconciliation workflows

Reconciliation is a long-standing core workflow in OpenRefine. Open GitHub issues show the need for clearer error handling, better user feedback, and more predictable behavior when reconciliation depends on external services. This work will address limitations that affect the user experience of reconciliation, including unclear failures, confusing match feedback, and reliance on workarounds for common matching scenarios. Strengthening this foundation will make reconciliation easier to diagnose, support, and maintain, while preparing the existing architecture for the local reconciliation work described below.

2. Align OpenRefine with the reconciliation API specification

OpenRefine will be updated to support the next version of the reconciliation API specification developed by the W3C Entity Reconciliation Community Group. This includes implementing the negotiation protocol and adding support for the draft v1.0 specification, as described in OpenRefine issue #7186. This work ensures compatibility with existing services while preparing the reconciliation architecture for more flexible local and external services.

3. Enable local reconciliation infrastructure

Building on the previous objectives, the project will introduce a native reconciliation capability within OpenRefine, enabling users to perform entity matching against local datasets without relying on external services or separate infrastructure. This includes designing and implementing a reconciliation endpoint that can operate directly on OpenRefine projects or locally defined authority files. Core services, including matching, preview, data extension, and suggestion, will be implemented with attention to performance, scalability, and configurability.

This work supports established use cases in data journalism, Open Source Intelligence (OSINT), and civic technology, where OpenRefine is used to prepare and reconcile high-stakes datasets. By improving the existing reconciliation foundation and enabling local reconciliation, the project lowers technical barriers, improves interoperability, and supports privacy-preserving data workflows used in public-interest contexts.

All development will be conducted in the open and integrated into OpenRefine's standard release cycle.

3.2 Briefly summarize open source licensing, community/contributor engagement, governance, status of documentation, etc. as relevant. Provide links to open repositories and documentation. [200 word limit]

Please note that you will have an opportunity to expand on these details if invited to a full proposal. This question serves as an overview and space to provide relevant links.

OpenRefine is an open-source project released under the BSD 3-Clause License, with development conducted publicly on GitHub: https://github.com/OpenRefine/OpenRefine. The project has been actively maintained for over a decade, with regular releases, open issue tracking, and public code review.

Governance is structured through a Core Developers Group, responsible for technical decisions, and an Advisory Committee, providing oversight and support. Decisions, processes, and roles are documented publicly (see https://github.com/OpenRefine/OpenRefine/blob/master/GOVERNANCE.md), ensuring transparency and enabling distributed participation. The project is maintained by a small group of core contributors, supported by a broader community of developers, documentation writers, and institutional partners.

Contributor engagement is supported through open collaboration channels, including GitHub discussions, a public forum (https://forum.openrefine.org), and regular community calls.

OpenRefine maintains extensive user and technical documentation (https://openrefine.org/docs/technical-reference/contributing), including manuals and developer

references. Documentation is updated alongside development and supports both users and contributors in adopting and extending the software.

3.3 Would your project benefit from additional non-financial support, such as OTF's Security Lab, User Experience and Discovery Lab, Impact and Engagement Lab, or another type of coaching/mentorship to support open sourcing and/or long-term sustainability efforts? Please briefly explain. [200 word limit]

Yes. This project would benefit from targeted support from OTF's User Experience and Discovery Lab, particularly in improving the usability and transparency of reconciliation workflows. Reconciliation involves semi-automated matching processes where users need to understand match quality, resolve ambiguities, and diagnose failures. Limitations in feedback and error visibility can make it difficult to understand where a reconciliation process has failed or how to interpret results.

We have discussed the project with Superbloom, an OTF-authorized UX Lab provider, and they have expressed interest in supporting this work if the application moves forward. Their involvement could help ensure that improvements to the underlying infrastructure translate into clearer, more reliable user interactions.

The project could also benefit from engagement support to better document and communicate reconciliation workflows across different communities, including journalism, civic tech, and Open Source Intelligence (OSINT) practitioners. These users often operate in high-stakes environments where data quality and reproducibility are critical. Strengthening documentation and knowledge sharing would support broader adoption and more effective use of the tool in these contexts.

3.4 Briefly explain your role/connection to this project, and/or why you are applying on behalf of this project, and who else may be involved in this work. [200 word limit] *

I am applying on behalf of OpenRefine as a contributor and community manager. I have been involved in the project since 2012, initially through documentation and training, and more recently focusing on coordination, governance, and sustainability. Since 2023, I have taken on a paid role to support community activities, including contributor onboarding, issue triage, release coordination, and facilitating communication between maintainers, contributors, and users.

My role in this project is to coordinate the work, support prioritization, and ensure alignment between technical development and community needs. This includes managing the grant, supporting collaboration, and integrating outcomes into the project's roadmap and release cycle.

The technical work will be carried out by experienced OpenRefine contributors, including members of the existing contributor and committer community with prior experience in reconciliation workflows. We have already identified interested individuals and will finalize allocation based on availability and scope as we prepare the final application.

This structure reflects OpenRefine's distributed model, combining focused development work with coordination and community input.

3.5 Please upload any relevant attachments.

4. I acknowledge:

- ☐ I am at least 18 years of age*
- ☐ My application will be dismissed if it does not fit within OTF's mission, values, principles statements.*
- ☐ I have read and understand OTF's Terms and Privacy policy.*
- ☐ I understand that all intellectual property created with support for this application must be openly licensed.*