OpenRefine

Contact: Martin Magdinier - Project Manager

martin@openrefine.org

Fiscal sponsor: Code for Science & Society, a 501(c)(3) public charity

🌍 openrefine.org

🧭 Mission & Vision

📊 2025 Report

📖 Usage & Impact

🚀 Roadmap

📜 Governance

Advancing Responsible AI Integration in Open Source Collaboration

OpenRefine & Alfred P. Sloan Foundation

As artificial intelligence (AI) accelerates innovation, open-source communities face new challenges and opportunities in adapting their practices to support AI-enabled workflows. OpenRefine is uniquely positioned to help researchers explore this frontier and serve as a model for other open-source projects grappling with their relationships to AI-generated code. OpenRefine's low-code interface also offers a natural entry point for domain experts to collaborate with AI models at scale, without needing advanced programming skills. Because of the diversity of researchers who interact with our software, as well as our robust communication channels, we have seen a visible shift towards AI adoption within our community, where contributors are generating useful features for OpenRefine by directing AI agents and leveraging other AI tools through the application itself.

Our positioning to support the transformation of research processes is enabled by OpenRefine's open-source ecosystem, which fosters rich community engagement and an effective contributor experience, enabling distributed collaboration and open innovation. In the last 12 months, we have seen examples of how our contribution experience, combined with clear, community-driven prioritization, has led to innovative new AI and LLM use cases with OpenRefine.

However, our open source community has further work to do to develop a shared understanding of how to evolve its processes to support AI-assisted contributions and AI-focused tool development. Providing better support and guidance to developers making AI-assisted contributions will broaden the possibilities for customizing OpenRefine to cater to our diverse

community needs. We believe that non-developer users in our community would also benefit from documentation about how to use AI tools within the OpenRefine application to improve and enhance their existing workflows. We seek to explore this topic by engaging our large and diverse community in focused and intentional ways, documenting our work, and sharing openly to benefit other open source software communities.

With Sloan's support, we will identify, assess, and implement ways to responsibly incorporate AI in the open-source contribution process and in data exploration and transformation workflows for non-developers.

User-Led Innovation and Open-Source Contribution

Over the last 12 months, we have seen an emergence of AI-focused extensions and tooling, as well as an increase in AI-assisted user solutions and contributions. One striking example occurred when a post on OpenRefine's forum gained substantial traction in October 2024. What started as a brief description of the user's attempt to connect OpenRefine to an LLM to run the same prompt on each line of a large dataset (10k items) turned into an exchange that engaged a wider group of contributors. Automating this process without OpenRefine would have required advanced computer programming skills. In early 2025, the community prioritized this work, and a code contributor converted the experiment into a ready-to-use, plug-and-play extension. The solution improved over the following months, with over 11 users and contributors collaborating on the extension and many more using it.

This is one of the many ways in which OpenRefine's community framework (detailed below) enables its members to express and document their needs, leading to the allocation of development resources and resulting in lasting, high-impact improvements of OpenRefine's software.

Thanks to the CZI EOSS program, we have developed and institutionalized a durable community framework that balances the needs of software users and code contributors. These practices have significantly improved OpenRefine's responsiveness, innovation capacity, and contributor retention. Key initiatives include:

- A moderated forum that enables meaningful discussion, technical support, and cross-domain collaboration
- An annual community event (BarCamp) to deepen relationships and foster innovation
- Regular user surveys to document use cases
- The creation of a goalpost list summarizing community requests
- Mentorship and onboarding pathways for early career developers through programs like Google Summer of Code and Outreachy

This work is made possible by focusing our resources on two key community support roles:

- **Developer Success Lead:** This role provides first-line technical support, answers contributor questions, and ensures that bug reports, pull requests, and releases are

reviewed and managed in a timely manner. These behind-the-scenes tasks ensure continuity, reduce contributor friction, and maintain the project's operational integrity.

- **Project Manager:** Acting as the connective tissue across community and institutional partners, this role documents use cases, gathers feedback on priorities, coordinates events, and stewards long-term relationships. The role is critical for ensuring that OpenRefine's roadmap remains aligned with the evolving needs of researchers, funders, librarians, and data practitioners (including industry users).

Meeting the Moment: Open Source Processes and AI Generated Customization Within OpenRefine

Throughout 2025, we have noticed a sustained increase in AI-assisted contributions. This raises novel questions for OpenRefine's community. How should our contributor processes evolve to account for this shift? What norms, tools, and expectations are needed to support AI-enhanced contribution, without compromising software quality or community trust?

As a project looking to democratize data transformation for non-developer users, we see the tremendous potential of AI-facilitated open-source contributions, as well as a range of pitfalls. OpenRefine's software user and code contributor communities have historically had little overlap, as the users tend to be non-developers. With the rise of AI, we additionally see the potential to bridge this gap and allow domain experts to prompt their chosen LLM to customize OpenRefine to meet their specialized needs. We want to develop pathways for proposing, reviewing, and accepting (or declining) custom functionalities that would be useful to the larger community.

We have real-world examples of this type of workflow already. For example, in October 2025, a long-time user was able to generate new OpenRefine functions and package them into an extension using Claude AI, serving the needs of a data journalist and the OSINT (Open Source INTelligence) community.

With support from the Alfred P. Sloan Foundation, we will thoughtfully explore whether providing the right contributor framework can help a new type of domain-expert AI-assisted code contributor build extensions and customization, accelerating innovation and increasing OpenRefine's relevance to their communities. Specifically, we will test how to best adjust our community contribution structure and governance to effectively adapt to AI-generated contributions. Key questions include:

- What community norms and infrastructure are needed to support AI-assisted code contributions?
- How can we safeguard quality and trust, while lowering barriers for domain experts to meaningfully participate in the OpenRefine codebase?
- What technical artifacts (e.g. prompt templates, test frameworks) must be included in pull requests that have been generated in collaboration with AI agents?

- What documentation about the use of AI tools would be the most useful and actionable for non-developer users?

Project Scope, Duration, and Budget

We propose a 12-month project focused on developing a contributor and reviewer framework to responsibly integrate AI-generated contributions into OpenRefine. This work will expand our current governance and contributor-support model to ensure software quality, maintain trust, and uphold community values as AI-assisted contributions grow.

Community Guidelines for AI Contributions:

- Facilitate discussion and co-create guidance for responsibly accepting AI-generated contributions to both the main codebase and external extensions.
- Outcome*:* Publish an AI Contribution Playbook, along with updated contributor and reviewer guidelines, drawing on the CSCCE community playbook model (for which multiple team members have been trained).

Practical Testing via Internal Extension Development:

- Develop at least one AI-assisted extension internally, using existing community goalposts (e.g., "multi-key joins" or "book-style title casing"), to assess what tooling and safeguards are needed to support AI-assisted contributions.
- Outcome: Update technical artifacts such as our sample-extension repository, PR templates, and prompt guidelines. Establish review and guardrail policies for extensions created with minimal developer oversight.

Community Engagement and Knowledge Sharing:

- Host structured discussions through GitHub, our community forum, and virtual workshops. Organize an in-person BarCamp to test and socialize ideas.
- Outcome: Disseminate learnings through public governance updates and present our progress at one or more open-source conferences in 2027.

Exploration of Documentation Needs:

- Survey non-developer users who are engaging with AI tools through the OpenRefine interface to better understand their use cases.
- Develop a plan for how best to document the workflows most essential to this user group.
- Share these recommendations with the community.

Measurement and Reporting: Define metrics to track AI-assisted contributions across core and extension repositories, including pull request activity and extension development volume.

Budget: USD 150,000

- Developer Success Role: $75,000 (0.5 FTE) for technical community support, triage, reviews, extension testing, and experimenting with AI-assisted code review practices.
- Project Manager Role: $45,000 (0.5 FTE) to coordinate community feedback loops, facilitate documentation, organize workshops and events.
- BarCamp (in-person): $20,000
- Conference Dissemination: $10,000