Open Source for Science Fund

Open Source for the Life Sciences (OS4LS)

Letter of Intent Submission

Submission ID	OS4LS-2f3c47a2-ce57-40a4-98b2-baa6fdfefc85
Submission Date	2026-06-08T16:44:42.606Z

Section 1: Applicant Information

Question	Answer
First Name	Martin
Last Name	Magdinier
Email	martin@openrefine.org
Organizational or Institutional Affiliation	OpenRefine
Country of Residence	Canada
Have you previously received funding as a PI under the CZI EOSS program?	Yes
Organization Name	Code For Science and Society
Organization Website	https://www.codeforsociety.org/
Organization Country	United States
Organization Type	Fiscal sponsor

Section 2: Proposal Information

Question	Answer
Proposal Title	Reproducible AI curation in OpenRefine

Short Summary	OpenRefine is a free, open-source, local-first tool for cleaning, transforming, reconciling, enriching, and reviewing messy tabular data. It is used in the life sciences to prepare, organize, and improve the quality of research data. Its operation history lets users document, audit, and rerun data-wrangling steps across similar datasets. Building on OpenRefine's EOSS-5 reproducibility work, this proposal extends that model to AI-assisted curation. Life-science researchers are increasingly exploring large language models for data extraction, annotation, classification, and validation. These tasks are especially relevant to biomedical literature review, clinical text curation, biodiversity and specimen records, and other workflows in which researchers need to convert text-rich records into structured fields. Many current LLM workflows make it difficult to apply the same prompt across a dataset, review outputs row by row, or preserve the information needed to reproduce results. OpenRefine already has a community-contributed LLM extension. It lets users apply prompts across rows inside an OpenRefine project. This integrates hosted or local LLM models into the tabular curation workflow and provides an interactive environment for users to inspect inputs and outputs, filter errors, and revise prompts. The proposed work would make this LLM-assisted workflow more usable, reliable, and reproducible for life-science research. It would provide: * An onboarding wizard with embedded setup instructions to help researchers connect LLM services without manually assembling endpoint URLs, model names, and provider-specific settings. * Preconfigured paths for popular local model services and reusable provider templates (Ollama, LM Studio, etc.), enabling users and institutions to add commercial, institutional, sovereign, or self-hosted model services without requiring custom extension code for each service. * Support for token-efficient workflows, including prompt compression, structured outputs with tight schemas, and model-selection guidance based on the curation task and available models. * Large-scale tabular AI workflow support by improving reliability for long-running and large tabular operations through improved progress and error handling, including support for batch APIs * Integration of AI-assisted operations into OpenRefine's reproducibility model by recording prompts, model parameters, response format, and related execution metadata in OpenRefine's operation history and exported workflows. The result would give researchers a human-in-the-loop interface for AI-assisted data curation within a tool already used in research and life-science workflows. As the use of AI in academia is increasingly discussed and debated, this project would support AI-assisted annotation and information extraction across tabular datasets in ways that are auditable, verifiable, documented, and reproducible, while preserving the transparency expected in research workflows.

Expected Value	The project will make LLM-assisted annotation, classification, and extraction reproducible in OpenRefine. These are emerging uses of LLMs in life-science research. In biomedical literature review, researchers could import article metadata, abstracts, screening notes, or extracted passages into OpenRefine, then use an LLM to classify papers against inclusion criteria, extract keywords or values, or flag whether a paper supports or contradicts a research claim. Outputs would appear as reviewable columns that researchers can filter, inspect, correct, and compare against the source text and existing metadata. This pattern also applies to scientific or clinical text curation for extracting structured attributes from sample labels and notes, including identifiers, locations, and specimen attributes, as well as harmonization workflows that classify variables or prepare candidate terms for review against authoritative vocabularies. In each case, the LLM output is not the final assessment; it is an iterative curation step that domain experts need to review, track, and document. OpenRefine would provide a guided interface for AI-assisted workflows, benefiting domain experts without requiring custom scripts or dedicated software support. Researchers could apply classification and extraction at scale while keeping outputs reviewable and methods documentation reusable. This would extend the reproducibility work developed through EOSS-5 to current AI-enabled life-science data curation.

Landscape Analysis	Researchers who want to apply AI to tabular data currently choose among several tool categories: * Spreadsheet software is common in research because it is familiar and convenient, but it offers limited support for tracking, auditing, and reusing data-wrangling steps. * Programming languages such as Python, R, MATLAB, and SAS can leverage LLM-specific packages such as LangChain and LlamaIndex, but they have a steep learning curve and lack an interactive tabular interface for reviewing edits. * Chat-based tools make LLMs accessible to researchers, but they are not designed to support reproducible structured data curation. * Emerging no-code batch LLM tools like BatchGPT often support row-wise prompting, but lack data curation features and robust support for scientific reproducibility. * Spreadsheet-like AI tools such as Hugging Face AI Sheets or OpenClay show growing demand for applying LLM prompts across tabular data. * Enterprise workflow platforms such as Dataiku and self-hosted automation platforms like n8n can support governed AI workflows, but they are heavier systems. OpenRefine already supports a wide range of scientific communities, including social, natural, and health sciences. Created more than 10 years ago, it now counts over 21,500 downloads per month and about 1,100 academic citations per year. The grant would enhance the LLM extension to integrate AI models in research workflows, ensuring usability, reliability, and reproducibility for current and new users.
Project Software Name	OpenRefine, OpenRefine LLM Extension
Project Software Repository URL	https://github.com/OpenRefine/OpenRefine, https://github.com/sunilnatraj/llm-extension
Project Software Website URL	https://openrefine.org/,

Categories	Knowledge representation and ontologies, Workflows and computational pipelines, Interoperability
Funding Track	Track 1: Domain-specific Tools
Statement of PI Involvement	Confirm
Policies	Confirm

Thank you for your submission. Please reach out to info@os4science.org with any questions and reference your submission ID as noted above in the email subject.