Angles

https://os4science.org/funding_opportunity/os4ls/ due June 8, 2026

Track 1 project

Different angles (which can be combined):

1) Federated repository curation workflow

This angle positions OpenRefine as a contributor-facing curation layer for generalist repositories, supporting

- metadata round-trips (Can we make something generic that needs to be adapted with a connector to a different provider?),
- validation-report-guided cleaning based on frictionless data,
- reusable curation workflows.

Questions:

- How do we manage the diversity of repository format and governance
- Is it safe to apply without a clear partner?
- Should we focus on general support for standards like frictionless and FAIR data?
  - DataCite, Dublin Core, OAI-PMH, schema.org, DCAT/DDI where relevant, Frictionless validation reports, and controlled vocabularies.
- If so, should those frameworks be widely used within the Life Science ecosystem?

It is grounded in light contacts with COS/OSF, Dryad, Vivli, Digital Science, and the broader GREI context

- https://forum.openrefine.org/t/discussion-how-openrefine-can-support-federated-data-repositories-for-open-science/2146
- 20251205 Dryad - Dryad is a particularly relevant example because it uses Frictionless Data to provide feedback on uploaded tabular files; OpenRefine could help contributors act on those reports through guided cleaning and export of corrected data.
- 20251204 Center for Open Science (COS / OSF)
- Julie has a connection with https://www.frdr-dfdr.ca/repo/ - FRDR is another useful reference point because its FAIR documentation connects repository deposits to DataCite Canada, Dublin Core, DataCite XML, OAI-PMH, custom metadata mappings, and controlled vocabularies such as FAST (https://www.frdr-dfdr.ca/docs/en/fair_principles/) .

https://app.diagrams.net/?splash=0#G11bPZyvCQt4i_CWQQsjj-d8ABdNlXxWYO#%7B%22pageId%22%3A%22Ay0KQnBLcHLrCbcQKrLX%22%7D

Ask for feedback from

- Hbz -> Dublin Core
- COS
- Vivli
- Digital Science -> Cat Allman
- Dryad
- Check the CRM

What needs to happen?

- Mapping
- Validating
- Generating the files in the right format

2) Life-science reconciliation and validation workflows.

Develop reference workflows and tests around concrete life-science entities: drugs, diseases, organisms, publications, authors, institutions, clinical variables, and controlled vocabularies.

- OpenAlex reconciliation (spoke with Jason Priem during the CZI Event, he was e not super passionate about since they focus on automation and not human in the loop)
- There is already infrastructure but we can strengthen it
- Align with Native Reconciliation with arbitrary external datasets goalpost for which we already submitted (and lost) several grant applications. OTF is application is also on that topic.

3) AI-assisted curation workflows inside OpenRefine

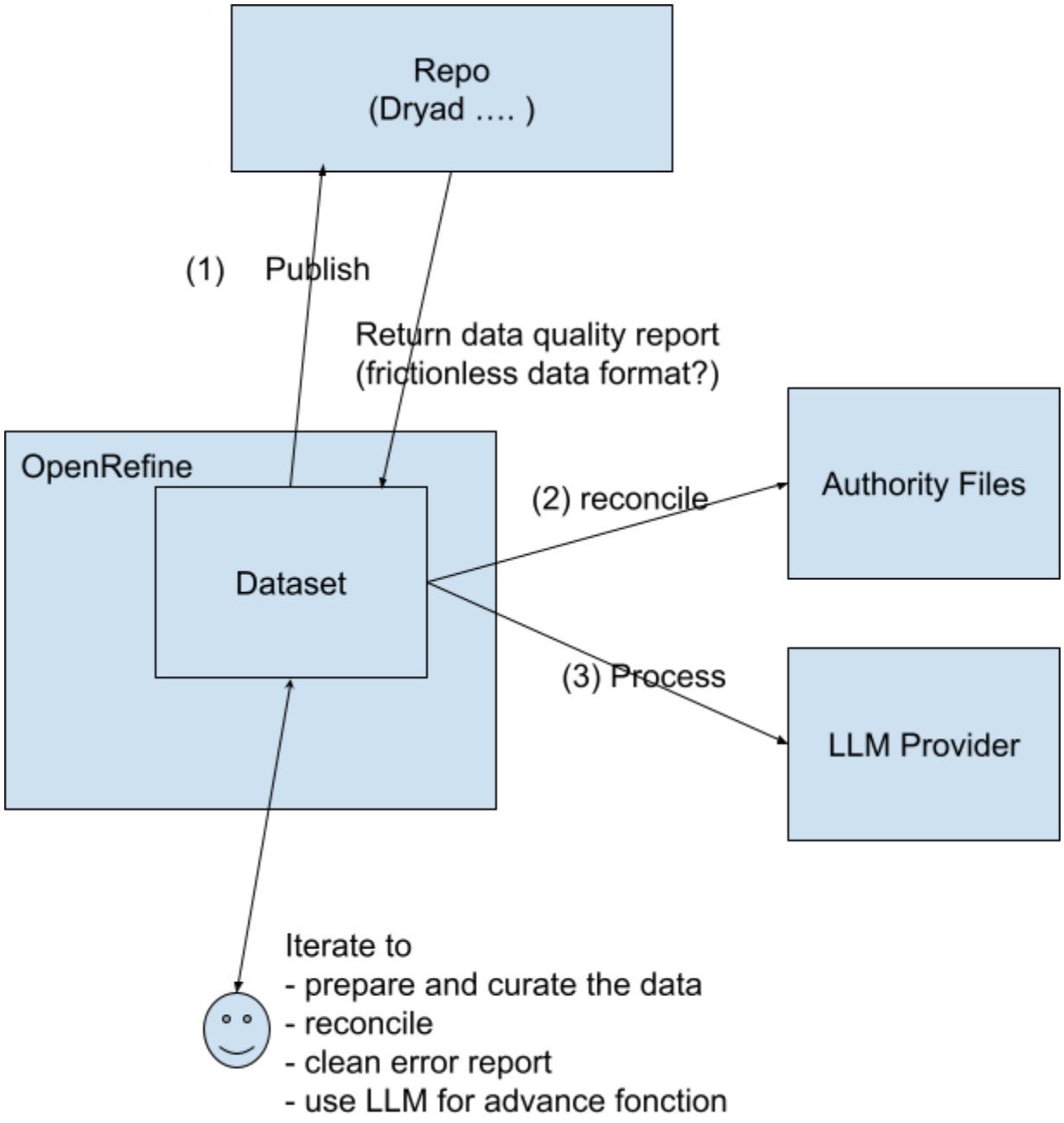
This builds on the existing `llm-extension` as a first step toward AI-assisted cleaning, enrichment, classification, and transformation. The stronger version would add traceability, prompt history (This is something Julie anticipates growing as a requirement in academia), validation, local/private model options, and clinician or domain-expert review.

Updating to support more provider ?
Address some of the gap in the documentation
Make it 0 config set up
Add a reasoning column for the AI to be able to explain its reasoning (do we need other traceability output)?

4) Frictionless report

Support end-to-end pipeline

- Curation
- Normalization
- Validation
- Linking with other datasets
- AI assisted for advance reasoning (selecting the reconciliation candidate)



-

🛑 Metadata Curration

Disclaimer	I used an LLM to help me generate this document.
Model used	Chatgpt 5.5
Document status	Exploratory for Review I am looking for comments and disagreements
My Knowledge on the topic	Advanced. Several studies have been done on the topic before drafting the proposal. Ideas are mine

Title of the proposal

OpenRefine for life-science metadata curation

Short summary (3000 characters)

OpenRefine is an open source tool widely used by researchers to clean, transform, enrich, and reconcile messy tabular data before analysis or publication. In an initial scan, we identified more than 30 publications and research outputs citing or documenting OpenRefine use in life-science and adjacent workflows, including bibliometric/scientometric analysis, pharmacovigilance and drug-name standardization, clinical and registry data harmonization, biodiversity data cleaning, and FAIR/semantic metadata preparation. These workflows share a common bottleneck: researchers manually map fields, normalize values, validate repository requirements, document transformations, and prepare metadata or data packages for reuse.

This proposal will add repository metadata curation workflows to OpenRefine, focused on preparing life-science data and metadata for reuse, repository deposit, and AI-ready downstream analysis. It will strengthen OpenRefine as a user-facing curation environment before repository deposit or reuse, producing validated metadata packages and change reports that can be reviewed or staged for ingest.

The proposed work is grounded in recurring user stories from life-science and adjacent research communities: bibliometric researchers cleaning publication metadata before sharing datasets; pharmacovigilance researchers documenting drug-name normalization decisions; clinical and registry data curators checking required metadata before deposit; biodiversity researchers validating collection metadata; and repository metadata specialists or data stewards preparing machine-readable metadata for discovery and reuse.

The proposed work has four connected components.

- First, an **OAI-PMH harvester** will allow users to create OpenRefine projects from repository metadata endpoints, with discovery of available sets and metadata formats, support for pagination/resumption tokens, and import of Dublin Core and DataCite records where available. This would complement existing OpenRefine project creation options, including creating projects from local files or sets of artifacts through the FilesExtension. Together, these input paths would support both users working from existing repository metadata and users starting from local research artifacts who need to generate, validate, and export repository-ready metadata for publication.
- Second, a **generalized metadata profile layer** will build on the interaction model already proven by the OpenRefine Wikibase extension: users map project columns to a target schema, validate the mapped data, and preserve the mapping as a reusable workflow configuration.
- Third, OpenRefine will **support DataCite and Dublin Core** as initial metadata profiles, with validation of required and recommended fields, identifiers, datatypes, repeatability, controlled values, and export structure.
- Fourth, a **repository profile configuration** module will let repositories or users load local application-profile constraints on top of the base DataCite or Dublin Core rules, reflecting the reality that repositories implement shared standards differently.

The result will be configurable, standards-aware curation components that researchers can chain with existing OpenRefine capabilities.

Out of scope, for now, are repository hosting, database operations, a new sharing platform, and direct publication connectors to each repository platform, because platform write-back depends on separate APIs, credentials, permissions, and governance models.

In life-science workflows, users often need to enrich metadata with external identifiers such as ORCID, ROR, DOIs, funder IDs, clinical trial identifiers, taxonomic identifiers, drug vocabularies, and disease vocabularies. OpenRefine already supports this kind of enrichment through APIs and reconciliation services, and the existing LLM extension lets users apply the same prompt across records using a commercial or locally hosted model. This proposal does not build new AI or reconciliation services; it adds the metadata profile and validation layer needed to make user-configured enrichment workflows easier to inspect, reuse, and validate.

Expected value (1500 characters)

Success means that life-science researchers and data curators can use OpenRefine to prepare repository-ready metadata through a guided, standards-aware workflow rather than relying on ad hoc spreadsheets, scripts, and hand-written export templates.

For researchers, this would make common preparation tasks more reliable: bibliometric researchers could clean publication metadata before sharing datasets; pharmacovigilance researchers could document how drug-name variants were normalized; clinical and registry data curators could check required fields before deposit; biodiversity researchers could validate collection metadata; and repository metadata specialists, data stewards, and research software teams could produce machine-readable metadata with clearer provenance.

For repositories and downstream systems, the value is better-structured metadata before ingestion. Repository curators would receive validated DataCite or Dublin Core metadata packages, local profile checks, and change reports that make review easier. Discovery systems and analysis pipelines would benefit from more consistent identifiers, controlled values, and documented transformations.

For OpenRefine, the work would generalize the existing Wikibase-style schema mapping pattern into a broader metadata profile layer. This would create reusable infrastructure for future standards and repository profiles, rather than one-off import/export templates.

For AI enablement, the proposal supports the data-preparation layer needed before model training, retrieval, evaluation, and agentic analysis. This project does not build new AI models. Instead, it helps researchers produce structured, standards-aligned, validated metadata that can be reused in AI-enabled workflows and large-scale data analysis.

Landscape analysis (1500 characters)

Life-science researchers and repository users currently rely on a mixture of spreadsheets, repository web forms, custom scripts, ETL tools, bibliometric tools, repository APIs, and standards-specific templates to prepare data and metadata. Related tools include Excel, Google Sheets, Python/R scripts, VOSviewer, Bibliometrix, Dataverse tooling, repository deposit interfaces, CEDAR, and metadata export systems in platforms such as Dryad, OSF, Zenodo, Figshare, Vivli, Dataverse, Mendeley Data, and FRDR. Prior discussions with repository stakeholders, including COS/OSF and Dryad, point to recurring needs around metadata quality, bulk review, validation feedback, and pre-ingest curation. FRDR and other repository examples also show that standards such as DataCite, Dublin Core, and OAI-PMH are already part of the repository interoperability landscape that this work would connect to.

OpenRefine occupies a different position in this landscape. It is a mature open-source desktop tool designed for interactive, reproducible cleaning, transformation, reconciliation, and enrichment of messy tabular data. It is already used in life-science-related publications for bibliometric analysis, pharmacovigilance, clinical and registry data preparation, biodiversity workflows, and FAIR/semantic data work. Its user base includes researchers, librarians, data curators, digital humanities practitioners, open-data teams, and repository-adjacent communities.

Compared with spreadsheets, OpenRefine offers stronger transformation history, clustering, reconciliation, and API-based enrichment. Compared with custom scripts, it is more accessible to non-programmers and domain curators. Compared with repository deposit forms, it supports bulk cleanup and reusable workflows before deposit. Compared with metadata-specific tools, it is more flexible for messy real-world tabular data.

OpenRefine is already used upstream of AI and data-intensive workflows because it helps prepare structured input data. The proposed work would make this role more explicit by adding standards-aware metadata mapping, validation, and export for repository-oriented life-science workflows. It would also better anchor OpenRefine within the repository and research data management ecosystem by creating clearer integration points with platforms and standards already used by its community, including OAI-PMH, DataCite, Dublin Core, repository APIs, local application profiles, and external enrichment services. OS4LS prioritizes tools that curate and structure scientific data for model training, AI workflows, and interoperability frameworks; this proposal targets that preparation layer directly.

—-

3 parts

- Explain the user story
    - Literature review
    - Artifact publishing
    - AI angles - see how Dublin Core is pitching it https://www.dublincore.org/
- Explain
    - what OpenRefine already support
    - What we will add
- What we will do in details
    -

Communications with FRDR

Slack message from Neha Milan

Thank you for thinking of FRDR. This is an exciting opportunity, and point 1 aligns well with where we're heading. With the new Tri-Agency data deposit policy landing in 2026, depositor-side preprocessing that lifts metadata and data quality upstream before coming to our curation team for review is real leverage for us.

Quick answers to your technical questions first. Globus is required only for our large deposits (>5GB or >300 files); smaller deposits come through our web interface, and our experience with smaller datasets is solid (we've a lot) even if they aren't the primary use case. On interaction beyond Globus more broadly — for the kind of metadata round-trips point 1 contemplates, the specific API surface you'd need is something I'd want to scope together with one of our engineers if we decide to proceed with this.

Before committing further, I'd like to better understand a few design constraints that are important for FRDR. Based on the limited context shared so far, I have a few questions. Many of these likely converge toward a single capability: a configurable mode in OpenRefine that restricts or disables outbound network calls (e.g., reconciliation, validation, LLM integrations) and supports locally cached authority files. If that is already within scope, several of the questions below may naturally resolve themselves:

- Embargo handling: Curators routinely work with depositors on embargoed datasets, and external reconciliation calls during that process could introduce pre-publication leakage risks. The same concern would also apply to our sensitive data pilot.
- LLM integration (point 3): External providers would not be suitable for embargoed and sensitive workflows. Is local or self-hosted LLM support being considered?
- Bilingual support: We operate in both official languages. Is reconciliation (and review) against French-language authority files within scope?
- Data sovereignty: FRDR is committed to Canadian-hosted data and aims to support Indigenous depositors operating within their own governance frameworks (e.g., OCAP/CARE). Default reconciliation endpoints could potentially create data egress pathways to non-Canadian infrastructure.

If you are able to share a draft of the point 1 section, I'd be happy to review it with these considerations in mind and continue the discussion from there. Once I have more information, Lee and I will discuss this internally and follow up. Happy to set up a call (after mid-June) as well if that would be useful.

Form

Information about the **applicant** and **host organization**

Title of the proposal

60 characters maximum

Short summary

Briefly describe the purpose of the proposal and the software project(s) it involves. 3,000 characters maximum

Expected value

If the proposal is successfully funded, what does success look like? We're seeking to understand: what type of capabilities the proposal is unlocking for the scientific community; how upstream and downstream software will be improved by the proposal; how the proposed work supports or implements novel functionality for AI enablement and large-scale data analysis. (1,500 characters maximum)

Landscape analysis

We are looking for proposals from software projects with demonstrated traction and adoption. Briefly describe other software tools that the audience for this proposal primarily uses (including proprietary alternatives, if they exist), and how the software project(s) in your proposal compare in terms of user base, adoption, functionality, and maturity relative to their target audience. You can add indicators of adoption and usage as needed. Please indicate if the software is used in AI applications and workflows. (1,500 characters maximum)

List of software projects

In this section you can list up to five open source software projects that will participate in this proposal. You can add projects using the “+Create” button, and include links to their repositories.

- **Software Project Name** * 30 characters maximum
- **Software Project's Repository URL**
- **Software Project's Website URL**

Categories *

- ☐ Data formats and storage
- ☐ Knowledge representation and ontologies
- ☐ Scientific computing
- ☐ Statistical modeling
- ☐ Workflows and computational pipelines
- ☐ Data visualization
- ☐ Interoperability
- ☐ Software ecosystem infrastructure
- ☐ Hardware acceleration and scalability
- ☐ Machine learning frameworks
- ☐ Agentic frameworks
- ☐ Benchmarking and evaluation tools
- ☐ Genomics and transcriptomics
- ☐ Structural and molecular biology
- ☐ Cell and developmental biology
- ☐ Evolutionary biology
- ☐ Immunology
- ☐ Biological and biomedical imaging
- ☐ Bioinformatics and computational biology
- ☐ Synthetic biology
- ☐ Spatial biology
- ☐ Neuroscience
- ☐ Infectious disease and epidemiology
- ☐ Computational drug discovery
- ☐ Other

Funding track:

- ☐ **Track 1 (Domain Specific Tools)**
- ☐ Track 2 (Foundational Libraries and Ecosystem Initiatives)

Applicant’s involvement in the proposal *

As an applicant, I confirm that I am a maintainer OR a designated representative of the software project(s) named in the proposal, AND that the proposed work is aligned with the projects’ respective roadmap and has support from their core maintainer community.

Confirm Applicant’s involvement in the proposal:

Policies *

I confirm that I understand and accept the Confidentiality Policy, Release of Liability, and the sharing of my proposal and reviews with current and future funding partners of the Open Source for Science Fund.

Notes from the QA Webinar

What we are looking for (1/2)

Proposals that target technical advances and address significant bottlenecks in OSS for science
Open source software projects enabling large-scale data analysis and AI-driven scientific applications and workflows

Priority areas (across both tracks) include

- Data curation, structuring & management for model training
- Benchmarks, standards & protocols for agentic workflows
- Scalability & hardware acceleration support
- Interoperability frameworks for AI-driven pipelines

AI is not required - but they are seeing a growing demand from the community to enable AI capabilities. How we support the new use cases

What we are looking for (2/2)

Proposals that target technical advances and address significant bottlenecks in OSS for science
Open source software projects enabling large-scale data analysis and AI-driven scientific applications and workflows

Likely unsuccessful or out of scope

- Early-stage prototypes with limited adoption
- Projects focused on a narrow target audience or niche use cases
- Proposed AI-assisted code rewrites without existing traction
- Proprietary software or freeware seeking funding to become open source
- Infrastructure for hosting data repositories, databases, or other types of scholarly outputs
- General maintenance without tangible deliverables
- Development of AI/ML models themselves
- Proposals focused on generating new datasets

Need broad audience > niche
Same maintainer community can rewrite their code via AI using this grant. But it needs to be the maintainer.

Q: Would tooling that prepares, validates, and exports metadata before repository deposit be considered in scope, as distinct from repository hosting or repository infrastructure?
> LOI can stay high level

> full proposal they expect technical depth on the science and technology side. They are reviewed by domain expert in the field + software engineering with open source background

A pure maintenance proposal is out of scope - but maintenance activities (documentation, community manager, backlog) are welcome as part of the grant application in trying to solve a significant bottleneck in the roadmap.

Q: Does the integration of software into an ecosystem of open projects qualify? Currently we are thinking about funding this integration work, of otherwise disperse projects.
> yes it does as long it is open source and in life science

Q: Does the integration of software into an ecosystem of open projects qualify? Currently we are thinking about funding this integration work, of otherwise disperse projects.
> Prototype and new code base are out of scope. We need to demonstrate traction and tackle a significant bottleneck

Q: in your opening slide, you mentioned "shared analytical tools for funders" to track which projects are emerging, should be sunsetted, etc. That sounds like a tool / dataset that would be really useful for maintainers too (e.g., for analyzing their dependencies, choices of backends, etc). Could you open-source that tool/dataset?
> they want to partner / fund project that help funder and maintainer - collecting citation and downloads. Looking up for that data should be easy to do.

Q: Does the integration of software into an ecosystem of open projects qualify? Currently we are thinking about funding this integration work, of otherwise disperse projects.
> yes

Q: If our ecosystem requires that information be collected in a database, does that rule it out as something that hosts data repositories?
> If a software tool is tied a data resource and individual data store
> Is the asset the data itself -> then it is out of scope
> If this is about interface with data and not tied to any data resource -> that's in scope

Q: Would tooling that prepares, validates, and exports metadata before repository deposit be considered in scope, as distinct from repository hosting or repository infrastructure?
> Metadata looks out of scope to Dario. Calling metadata, annotation about a dataset before publication -> out of scope
> Other type of metadata,
> Curation publication of metadata is out of scope for this call.

The eligibility phase is just about meeting the criteria of the calls.

We can do multiple application for the same project / applicant as long as the scope of work is different.