Native In-App Reconciliation of Arbitrary Local Data Sets

This proposal addresses the goal post Native In-App Reconciliation of Arbitrary Local Data Sets. It would align closely with the fund's objective to democratize access to data and promote decentralization.

We can build on the existing grant application 2025 Research Software Maintenance Fund.

Contact information

Your name	Martin Magdinier
Email address	martin@openrefine.org
Phone number	+1-503-383-1430
Organisation	Code For Science and Society
Country	USA

General project information

Proposal name	Native In-App Reconciliation of Arbitrary Local Data Sets

Website / wiki	https://openrefine.org

Please be short and to the point in your answers; focus primarily on the what and how, not so much on the why. Add longer descriptions as attachments (see below). If English isn't your first language, don't worry — our reviewers don't care about spelling errors, only about great ideas. We apologise for the inconvenience of having to submit in English. On the up side, you can be as technical as you need to be (but you don't have to). Do stay concrete. Use plain text in your reply only, if you need any HTML to make your point please include this as attachment.

Abstract: Can you explain the whole project and its expected outcome(s). (1200 characters)

European Galleries, Libraries, Arts, and Museums (GLAM) organizations are in the process of digitizing their collections to make them easily discoverable, reusable, and accessible to the public while enhancing preservation of the Commons. This process involves standardizing collection records, which often requires expensive outside services and technical expertise that smaller organizations simply lack.

We aim to develop a free, user-friendly tool within OpenRefine (a widely used, open-source software for cleaning up messy data). This new feature will allow staff at any organization to link their Excel spreadsheets or CSV files with shared identifiers and taxonomy. We will provide video tutorials, instruction manuals, and hands-on training to key staff at two organizations.

By building a native reconciliation engine directly into OpenRefine, we're establishing a durable commons infrastructure that will serve the global library and cultural heritage community indefinitely, ensuring ongoing maintenance of the tool and training local champions who will spread knowledge throughout their networks, creating a multiplier effect that extends far beyond our initial funding period.

Have you been involved with projects or organisations relevant to this project before? And if so, can you tell us a bit about your contributions? (optional, help to determine we are the right person to take on this project 2500 characters)

This project is led by the OpenRefine core team and has received widespread community support.

OpenRefine (lead applicant) is fiscally sponsored by Code for Science and Society, a 501(c)(3) charitable organization in the USA. OpenRefine leads the development and sustainability of the software application, including maintenance planning, technical debt reduction, contributor onboarding, documentation improvements, maintainability, and community support. The project is led by Rory Sawyer, Developer and Contributor Engagement Lead, and Martin Magdinier, Project Manager.

We will partner with **Ura** (https://ura.design) for the UX and design milestone of the project. Ura Design is a Berlin-based studio supporting open, ethical, and accessible tools for science, technology, and human rights. They specialise in strategic brand systems, interdisciplinary storytelling, and digital infrastructure for mission-aligned organisations. Their team brings over a decade of experience designing for complexity, connecting scientific research, policy, and advocacy into coherent, trusted narratives. They have collaborated with global partners such as The Tor Project, the Open Technology Fund, Internews, GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit), Censored Planet (University of Michigan), Hessian.AI (Technical University of Darmstadt), Interalia, and Emergencity.

Partner Organizations, see attached letter of support

NFDI4Culture, the consortium for research data on material and immaterial cultural heritage within the German National Research Data Infrastructure, relies heavily on OpenRefine to establish a needs-based infrastructure that serves our community of interest, ranging from architecture, art history and musicology to theatre, dance, film and media studies. They have allocated staff time to pilot testing, user feedback sessions, and documentation review. They will also or co-host at least one webinar in Germany to disseminate project results.

SODa focuses on building key competencies across the collection-related research data lifecycle. This includes ensuring data quality and compatibility through the use of standards such as persistent identifiers, linked data, and graph-based approaches, as well as enabling the enrichment and contextualisation of collection data through reconciliation with authoritative and localised vocabularies.

We also received a letter of support from **LaOficina Producciones Culturales** regarding their usage of OpenRefine.

We expect other European and non-European institutions to engage with the project as it was identified as a top feature request by the OpenRefine community.

Requested support

Explain what the requested budget will be used for?

Does the project have other funding sources, both past and present?

(If you want, you can in addition attach a budget at the bottom of the form)

Explain hardware, labor (including rate), travel cost , technical meeting - max 2500 characters

Team Organization

- **Martin Magdinier:** Project Management: hourly rate EUR 50
- **Rory Sawyer:** Development: hourly rate EUR 65
- **Ura Design** (hourly rate EUR 65) UX research & wireframes,
- **Code for Science and Society** fiscal sponsorship fees (15% of total grant)
- **Travel and Event:** EUR 3,000

Hardware & Infrastructure: No dedicated hardware purchase is required: all development and testing will leverage cloud-based CI environments and existing OpenRefine infrastructure servers.

Funding Sources

Present: There are no concurrent cash co-funders for this NLNet application; however, we will leverage the established OpenRefine community roadmap and in-kind contributions (CI hosting, volunteer tester time) to maximize impact at minimal additional cost.

Previous funding sources for OpenRefine are available at https://openrefine.org/funding

Project Plan

We expect the project to be completed within 7 to 8 months, depending on the availability of the different organizations. We broke it down into five milestones. Milestone scope and details are provided in answer to the question “What are significant technical challenges you expect to solve during the project, if any?”

Milestone 1: Team Mobilization and Project Kick-Off

- Duration: 30 days
- Estimated Effort: 15h
- Budget EUR 900
- Team
    - Rory Sawyer
    - Martin Magdinier
    - Ura Design
    - NFDI4Culture
    - SODa

Milestone 2.1 Technical Assessment

- Duration: 1 month
- Estimated Effort: 30h
- Budget EUR 2,600
- Travel and Event Cost
- Team: Rory Sawyer

Milestone 2.2 UX Design:

- Duration: 2 months
- Estimated Effort: 111h
- Budget EUR 7,005
- Travel and Event Cost:
- Team
    - Martin Magdinier
    - Ura Design
    - NFDI4Culture
    - SODa

Milestone 3.1: Implementation of the negotiation protocol

- Duration: 1 month
- Estimated Effort: 40h
- Budget EUR 2,600
- Team: Rory Sawyer

Milestone 3.2: Implementing support for the reconciliation API version 1.0 draft

- Duration: 1 month
- Estimated Effort: 174h
- Budget EUR 11,000
- Team:
    - Rory Sawyer
    - Martin Magdinier

Milestone 4.1: Determine the implementation architecture

- Duration: 1 month
- Estimated Effort: 24h
- Budget EUR 1,350
- Team:
    - Rory Sawyer
    - Martin Magdinier

Milestone 4.2: Match Service or Reconciliation Queries

- Duration: 1 month
- Estimated Effort: 80h
- Budget EUR 5,200
- Team: Rory Sawyer

Milestone 4.3 Preview Service

- Duration: 1 month
- Estimated Effort: 40h
- Budget EUR 2,600
- Team: Rory Sawyer

Milestone 4.4 Data Extension Service

- Duration: 1 month
- Estimated Effort: 40h
- Budget EUR 2,600
- Team: Rory Sawyer

Milestone 4.5 Suggest Service

- Duration: 1 month
- Estimated Effort: 40h
- Budget EUR 2,600
- Team: Rory Sawyer

Milestone 5.1: Training and Workshop - Reconciliation API (v1.0 draft)

- Duration: 1 month
- Estimated Effort: 23h
- Budget EUR 1,270
- Team
    - Rory Sawyer
    - Martin Magdinier
    - NFDI4Culture
    - SODa

Milestone 5.2: Training and Workshop - Reconciliation within OpenRefine

- Duration: 1 month
- Estimated Effort: 23h
- Budget EUR 1,270
- Team
    - Rory Sawyer
    - Martin Magdinier
    - NFDI4Culture
    - SODa

Compare your own project with existing or historical efforts. (4000 characters)

Data cleaning, transformation, and reconciliation tools can be categorized as follows:

- **1. Spreadsheet software** provides an entry-level interface to data manipulation but offers only basic functionalities and does not scale for fuzzy matching or support reconciliation processes.

- **2. Programming languages** like Python and R offer flexibility and reproducibility but have a steep learning curve.

- **3. Reconciliation clients** supporting reconciliation in a specific context:
    - **Alma-refine** is limited to a set of preconfigured reconciliation endpoints and is available only via the Ex Libris App Center
    - **Cocoda** focused on managing and creating mapping between knowledge organization and not on reconciling collection data with an authority file.
    - **SemTUI** is specific to the exist-db ecosystem.
    - The **RDF Extension for OpenRefine** by DERI at NUI Galway includes reconciliation against any SPARQL endpoint or RDF dump file but does not support tabular authority files.

Existing Reconciliation Framework. Other reconciliation frameworks exist, but they have limitations. These limitations include the requirement for the user to have a working coding

development environment in Python (datasette-reconcile, csv-reconcile) or Java (conciliator, reconcile-csv); or knowledge of semantic web technologies like SPARQL (grefine-rdf-extension) or SKOS (skohub). Additionally, many of these third-party reconciliation frameworks are no longer actively maintained and are vulnerable

Our local OpenRefine endpoint fills the gap by providing a native reconciliation feature directly within OpenRefine. Our integration will be

- **Truly Native:** By embedding reconciliation directly into OpenRefine application as a plugin, we deliver a seamless, consistent interface for all users without another service to run.
- **Offline & Sovereign Workflows**: Users can reconcile against local dumps directly on their computers, protecting sensitive data and ensuring uninterrupted service even when public APIs are throttled.
- **Fully Customizable:** By letting users configure the matching and scoring algorithm, they maintain their agency and control over how the data is processed.
- **Sustainable Commons Stewardship:** Leveraging OpenRefine's existing BSD-3-Clause governance framework and community roadmap, our engine will be embedded within the OpenRefine ecosystem and receive ongoing maintenance, ensuring that this work remains a durable, shared resource rather than a one-off development.

What are significant technical **challenges** you expect to solve during the project, if any? (5000 characters)

The OpenRefine community has already scoped and described their need in a series of GitHub issues. We have organized our project into six milestones, as detailed in the budget section, with each milestone aimed at addressing a specific technical challenge we have identified.

The project is divided into five milestones.

Milestone 1: Team Mobilization and Project Kick-Off allocate one month to mobilize our team and partner. During this phase, we will confirm the project timeline and resource allocation.

Milestone 2: Design Phase

This phase will include two sub-milestones.

2.1 Technical Assessment: First, the development team will determine the effort and complexity to create the reconciliation service against an existing OpenRefine project (as requested in #2003) or against a new data source. Enabling the reconciliation of data from one OpenRefine project to another would address the need to perform a "fuzzy join" between the two projects.

2.2 UX Design: At the same time, in a second submilestone, we will conduct a user experience (UX)- focused audit to understand how European institutions utilize reconciliation workflows and the barriers they encounter. We will publish our findings and use them to guide architectural and design decisions. This effort is led by Ura and in collaboration with NFDI4Culture and SODa.

As a result of this research, the Ura will produce the wireframe and design of the interfaces to create and configure the reconciliation endpoint to

- Import a new authority file (if we cannot reconcile against an OpenRefine project).
- Configure which columns are available for reconciliation, and for each one, specify the matching algorithms and scoring methods used. It is crucial that users have the agency to customize how records are matched and scores to best fit their use case.
- Export and import configurations as a JSON object, allowing them to be shared with other users.
- Start and stop the reconciliation service.

The design would follow the OpenRefine design system (https://openrefine.org/docs/technical-reference/openRefine-design-system).

Milestone 3: Development: Compliance with the New Reconciliation API (v1.0 draft)

The work will be divided into two sub-milestones.

3.1 Implementation of the negotiation protocol: For OpenRefine to support reconciliation endpoint supporting different versions of the Reconciliation API (v0.1, v0.2, and v1.0 draft), we need to implement a negotiation protocol as described in https://github.com/reconciliation-api/specs/issues/78

3.2. Implementing support for the reconciliation API version 1.0 draft: The reconciliation feature in OpenRefine relies on a unified protocol (Reconciliation Service API v0.2 - https://www.w3.org/community/reports/reconciliation/CG-FINAL-specs-0.2-20230410/), which is used to communicate with various web services offering data matching functionalities for various data sources. The W3C Entity Reconciliation Community Group (https://www.w3.org/community/reconciliation/) has been working on a new version of this protocol (https://reconciliation-api.github.io/specs/draft/), addressing a number of problems in the current one. For instance, it is currently very cumbersome to reconcile entities in OpenRefine when one does not have a column with names of those entities, because the existing protocol assumes that those names must always be provided (#6044). We would like to add support to OpenRefine for the new protocol to address this issue, and various others (see below). We would like to maintain support for the existing protocol in OpenRefine, so that existing reconciliation services continue to work. As we do this work, we plan to bring up issues to the W3C group, so that the new specifications can be improved accordingly. This is a good

opportunity to update our current reconciliation code to utilize an external library (https://github.com/wetneb/ReconToolkit) instead.

Relevant issues: #7186 #6234, #6053, #6044, #4715, #3139, #2332, #2075

We are implementing this feature first, so it will allow the community to review and comment on the implementation while we work on Milestone 4.

Milestone 4: Development: Creation of the reconciliation endpoint

Based on the results of milestone 3 and the success in implementing support for the version 1.0 draft of the reconciliation API, we will decide if we develop the reconciliation endpoint following the 0.2 (which already has several examples of successful implementation) or 1.0 draft specification.

4.1 Determine the implementation architecture: First, we will determine if it is best to develop this feature as part of OpenRefine core or as an extension. We will first need to assess how an extension can access OpenRefine project data and how we can expose it through reconciliation.

Then, depending on the selected API version, we will break down our development into one submilestone per service defined in the Reconciliation API specification. Each service will be implemented as a backend functionality and supported by a test suite. Such an implementation would be useful even if not all the features of the reconciliation API are implemented initially.

4.2: Match Service or Reconciliation Queries, which will include

- Implementing matching and scoring methods based on algorithms already supported in OpenRefine.
- Implementing the UI defined during the design milestone
- Taking scalability and performance management into consideration by creating a queryable index and managing resource allocation.

4.3 Preview Service including

- View entities
- Preview entities

4.4 Data Extension Service including

- Extend data

4.5 Suggest Service including

- Suggest entities
- Suggest types
- Suggest properties

Relevant issues: #2003 #176 and #941

We plan to engage with our partner organization (NFDI4Culture and SODa) and the OpenRefine developer community through the development of each service to test the reconciliation endpoint and validate our progress.

Milestone 5: Training and Workshop

In partnership with NFDI4Culture and SODa we will organize webinars and training sessions to disseminate the results and produce relevant user and developer documentation.

Describe the **ecosystem** of the project, and how you will engage with relevant actors and promote the outcomes? (2500 characters)

Ecosystem

Our project is part of OpenRefine's vibrant ecosystem. In the last 12 months, OpenRefine averaged 15,500 downloads per month, with 22 active contributors during this time. A total of 163 issues were created, and 157 were closed via 212 pull requests. OpenRefine received about 800 academic citations per year.

The in-app reconciliation feature was the second most upvoted request in our 2024 feature ranking survey (https://forum.openrefine.org/t/results-from-the-feature-prioritization-survey-2024/1847), highlighting the relevance and impact of our work.

Engagement

For this grant, we have secured a partnership with NFDI4Culture and SODa in Germany, who will participate in pilot testing, user feedback sessions, and documentation review. We will also reach out to other GLAM organizations we interacted with during our yearly conference and bi-yearly user survey to request their participation.

Our project will be open by default. We will publish every aspect of user research, project planning, and development on GitHub and our community forums to make these materials useful and reusable for other organizations. We will involve the OpenRefine Core Dev Group through regular open calls, design critiques, and code reviews. This ensures early feedback and smooth integration into the existing release cycle.

We will host three milestone webinars at the end of Phases 2, 3, and 4 to showcase progress and gather feedback. The final webinar (Milestone 4) will feature live demos and a Q&A session, influencing improvements to the OpenRefine documentation. Additionally, we will present the project's results at our annual Barcamp conference.

We also budget in-person workshops in Europe, where our design and development will meet with NFDI4Culture and SODa staff to test workflows, refine usability, and record case-study content.

The outcomes of this grant, including the integration of native reconciliation services into OpenRefine, comprehensive documentation, training materials, and pilot case studies, will be incorporated into OpenRefine’s BSD-3-Clause governance model. The code will be part of OpenRefine's regular maintenance and release schedule. The OpenRefine community will be able to recruit local ambassadors from partner institutions to champion the use and adoption of OpenRefine, as well as the new in-app reconciliation feature. We’re confident the module will become a durable part of the Commons infrastructure for GLAM and library data projects.