Sustainable Reconciliation Infrastructure for Biodiversity and GLAM Workflows

Size of grant

- **Large award (up to 2 years / £500k)**

Proposal summary / scope of work 500 words

Reconciling collection data with authoritative vocabularies enables curators, librarians, and researchers to align records with shared identifiers (such as taxonomic names or authority files), giving them the ability to link and access allied data. However, deploying reconciliation services requires highly qualified personnel, burdening researchers with limited technical resources and expertise, including institutions under the Galleries, Libraries, Archives, and Museums (GLAM) category, and in biodiversity and taxonomy. UK institutions have used tools like OpenRefine (https://openrefine.org) and the Royal Botanic Gardens, Kew (RBGK) Reconciliation Framework (https://github.com/RBGKew/Reconciliation-and-Matching-Framework) to simplify the reconciliation, metadata management, and digitisation workflows for over a decade. It will play an increasingly important role within the DiSSCo-UK digitisation programme, which aims to digitise the UK natural science collections.

RBGK's and OpenRefine's popularity among UK researchers incentivised us to seek to modernize RBGK's reconciliation framework legacy code to align it with today's reconciliation and integration best practices. This will accelerate the digitisation process at small and under-resourced institutions by alleviating some of the technical burden placed on researchers. It will also enable OpenRefine to continue modularising its support for extensions and external reconciliation services. The project aligns with the OpenRefine community roadmap as it will enable users to perform reconciliation with external datasets (https://openrefine.org/docs/technical-reference/goal-posts#native-reconciliation-with-arbitrary-external-datasets and https://github.com/OpenRefine/OpenRefine/issues/2003).

We have defined a 5-step plan to improve the maintainability, usability, and interoperability of the reconciliation infrastructure:

1) Needs Assessment. Led by Dr. Gray and RBGK and in collaboration with King's Digital Lab and RBGE we will conduct a UX-focused audit to understand how UK researchers and institutions use reconciliation workflows and what barriers they face. We will publish our findings and use them to guide architectural and design decisions.

2) Delivery Strategy. Based on the audit and with the input of the OpenRefine Core Dev Group, we will determine the most sustainable and accessible way to deliver reconciliation capabilities to users. Current options include a third-party application (current state), an OpenRefine extension, or a full integration with OpenRefine.

3) Refactoring and Integration of RBGK's Reconciliation Framework. This will be modernised by aligning it with OpenRefine's plugin architecture and the latest reconciliation API standard (https://reconciliation-api.github.io/specs/1.0-draft/). This includes:

- Structural code improvements (modularisation, test coverage, scalability)
- Integration into OpenRefine as a reusable plugin or module, easing maintenance and community contribution (as defined in step 2)
- UX improvements based on research findings to allow end users to configure matching criteria.

4) Core OpenRefine Enhancements. Update OpenRefine to

- Fully implement the latest reconciliation protocol to provide more transparent and detailed matching scores (https://github.com/OpenRefine/OpenRefine/issues/7186).
- Improve OpenRefine's extension system to support the integration of the RBGK reconciliation framework and enhance the overall developer experience (addition of hooks and API endpoints, increase in test coverage, and documentation improvements).

5) Documentation and Training. We will produce documentation, training materials (including video tutorials), and share updates through blog posts, case studies, and community outreach.

Categories of work

- **Technical**
- **Community**
- **Documentation**
- **Training**
- ~~Governance~~

Project Team 250 words

We comprise an international team with a proven track record of supporting the UK's research needs through collaborative and community-led software products.

Royal Botanic Gardens, Kew (Lead Applicant) will lead institutional engagement and refactoring of the reconciliation infrastructure. The Principal Investigator is Dr. Nicky Nicolson, a Fellow of the Software Sustainability Institute and the original developer of the RBGK Reconciliation and Matching Framework.

King's College London (Co-Lead) Dr. Jonathan Gray, will lead participatory design processes, usability testing, and UX-driven development in collaboration with staff at King's Digital Lab.

Royal Botanic Garden Edinburgh (RBGE) (Partner) Dr Mark Watson, co-chair of the World Flora Online (WFO), and Dr Elspeth Haston, lead of the herbarium digitisation programme, will align the development of the WFO reconciliation infrastructure with OpenRefine and develop training and documentation support in national digitisation programmes (DiSSCo-UK).

OpenRefine (International Co-Lead) 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 tool, 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.

The structure ensures the proposed work is maintainable, reusable, and driven by the needs of UK research institutions that rely on OpenRefine and reconciliation services.

Benefit to UK research

Briefly describe the expected value of the proposed work to UK research. You may wish to describe the fields of research or types of research method the software is used in, and describe the benefits to a particular community. You may also want to give examples of what research is enabled by the software. 250 words

OpenRefine is an open-source tool widely used across the UK's research ecosystem. It enables users without programming skills to install the software and perform advanced data cleaning, transformation, and reconciliation tasks that go beyond the capabilities of traditional spreadsheet tools.

We have documented usage at the British Library, the Natural History Museum, the National Archives, RBGK, and the RBGE. OpenRefine also supports digital culture research, including usage at the Centre for Digital Culture within King's College London's Digital Futures Institute.

In biodiversity and GLAM sectors, OpenRefine and its reconciliation framework are integral to the integration of information from separate sources (eg in the compilation of identification and conservation resources like Plants of the World Online and the World Flora Online) and the

digitisation of collections (e.g. in DiSSCo UK, where it will be used to support data capture, data cleaning, and data enhancement). As electronic data are increasingly attached to specimens from point of collection to downstream research use, data cleaning and reconciliation processes are crucial for usable data.

The software is adopted in training and practice:

- OpenRefine is included in the Carpentries Data Carpentry curriculum, with 171 OpenRefine workshops since 2015.
- OpenRefine was cited in over 174 academic publications in 2024, including 14 from the UK.
- The RBGK Reconciliation Framework is currently used by at least four institutions to provide reconciliation endpoints.

Landscape analysis

Briefly describe the other software (either proprietary or open source) that is primarily used by the research community addressed by the work in this proposal. Summarise, to the best of your knowledge, how the software in this proposal compares to these other software in terms of user base size, usage, and maturity. Describe, if appropriate, how the other software interacts with the software in this proposal. 250 words

Alternative to OpenRefine: 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 specific context:
    a. Alma-refine is limited to a set of preconfigured reconciliation endpoints and is available only via the Ex Libris App Center
    b. Cocoda focused on managing and creating mapping between knowledge organization and not on reconciling collection data with an authority file.
    c. SemTUI is specific to the exist-db ecosystem.
    d. The existing R package, Python library, and command-line interface make it difficult for non-programmers to use.

4. OpenRefine fills the gap between these categories. With a powerful graphical user interface, this category of software can be easily mastered by non-programmers, allowing them to work on both data cleaning and reconciliation tasks.

Alternative to RBGK 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.

This emphasises the need to support the development and maintenance of OpenRefine as a user-friendly gateway to accessing the RBGK framework and other comparable reconciliation endpoints.

Measure of impact (optional)

Briefly describe how you would measure the impact of the proposed work. You may have existing measures that you use, or you may propose new measures. For the Expression of Interest we are looking to understand how you would approach this, and what you think is important to measure. In the full application, you will be asked to define specific measures. 250 words

We are considering a number of measures to assess the impact of refactoring and integrating RBGK's Reconciliation Framework to OpenRefine:

- Increase in the number of institutions using OpenRefine in their digitisation process through the DiSSCo-UK programme.
- Number of additional researchers trained to use the refactored framework
- Number and origin of contributors maintaining the refactored reconciliation framework after 12 months.
- Number and origin of contributors maintaining OpenRefine after 12 months.