BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e134910
doi: 10.3897/biss.8.134910

OPEN ACCESS

Conference Abstract

Can We Standardize Name Reconciliaton via OpenRefine?

Dmitry Mozzherin‡, Deborah L Paul‡, Amanda Whitmire§

‡ University of Illinois, Champaign, United States of America
§ Stanford University, Pacific Grove, United States of America

Corresponding author: Dmitry Mozzherin (dmozzherin@gmail.com)

Received: 16 Aug 2024 | Published: 19 Aug 2024

Citation: Mozzherin D, Paul D, Whitmire A (2024) Can We Standardize Name Reconciliaton via OpenRefine? Biodiversity Information Science and Standards 8: e134910. https://doi.org/10.3897/biss.8.134910

Abstract

Scientific names in biodiversity represent one of the oldest identifiers used in science. As a result, a common repetitive task is being able to reconcile a list of scientific names against curated data sources. Reconciliation allows one to determine if names in a list are spelled correctly, whether they are currently accepted, and their nomenclatural status. There are several online and local resources that provide reconciliation services. We share here the potential in interoperability across reconciliation tools.

Global Names Verifier (GNverifier), Catalogue of Life, Global Biodiversity Information Facility (GBIF), Taxonomic Name Resolution Service (TNRS), LifeWatch, National Center for Biotechnology Information (NCBI), World Flora Online, Global Biotic Interactions (GloBI), Nomer, Wikidata, and others provide their own tools for name reconciliation. All these tools have their scope, design decisions, input, and output formats. It is often useful to do reconciliation using several such services, because they often include complementary data. However, with all the idiosyncrasies of services and lack of standardization, it is not an easy task (Islam et al. 2024). It would be great for researchers if all existing and future tools could be standardized. Then moving from one resource to another would be as easy as changing the URL. Implementing elements of Findable, Accessible, Interoperable, and Reusable (FAIR) data management principles would help to create such standards.

However, standardizing all existing and future resources to a common interface would be difficult. Some of them have no monetary or programmatic means to modify their code, while others have more urgent priorities. Some resources support a specific research path where adhering to a rigid standard might hinder their innovation. In this paper we suggest interoperability between reconciliation tools by implementing the OpenRefine Reconciliation Service. OpenRefine is a popular and powerful reconciliation and data cleaning application. It is used by many researchers for data transformation and normalization. Any service that implements the OpenRefine Service can be incorporated into data-management workflows just by providing the service's OpenRefine-compatible URL. Such compatible services can easily be discovered by providing their metadata in the OpenRefine Services Registry.

In this paper we discuss our implementation of the OpenRefine Service with the Global Names Verifier (GNverifier) reconciliation tool.

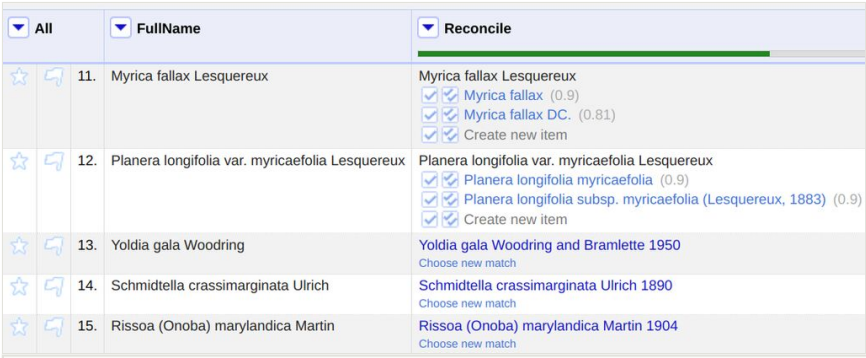
GNverifier is developed at the Species File Group as a part of the Global Names Architecture initiative. It offers a powerful, configurable, fast way to reconcile scientific names. GNverifier software aggregates data from more than 100 source datasets. Queries return currently accepted names when provided in a dataset. It allows finding matches for names that historically had several suffixes and can do fuzzy and partial matches. It sorts data by many factors to reliably provide the best available results. With a strong focus on software optimization and a sophisticated matching algorithm, it can process 2000 names a second, making it one of the fastest services available.

OpenRefine can use GNverifier directly because it is compatible with the OpenRefine protocol. As shown in Fig. 1, switching between GNverifier and Wikidata reconciliation of scientific names requires only change of a service URL.

Figure 1.

OpenRefine makes it easy to choose between several reconciliation services, in this case Wikidata and Global Names Verifier.

Implementation of the OpenRefine protocol might solve many standardization problems. Some resources already have it implemented (e.g., Wikidata, GNverifier Whitmire and Mozzherin 2023, WFO Plant List, IPNI). Many people already use OpenRefine for their other reconciliation needs. For them, the incorporation of name reconciliation would be especially beneficial because it will fit into their existing data-management workflow Fig. 2. Basic reconciliation by itself is standard by design and a big step forward.

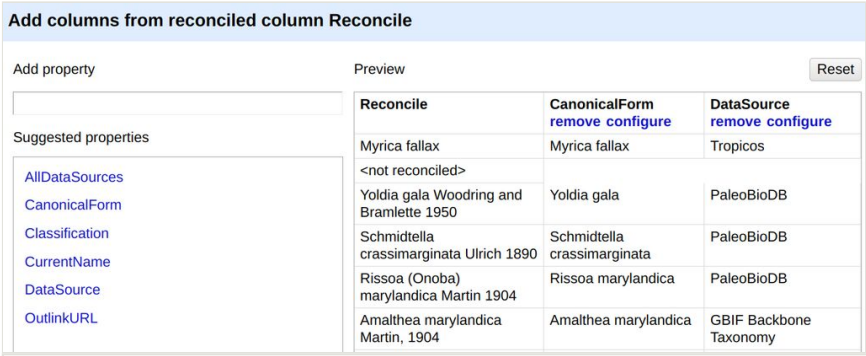


All	FullName	Reconcile
11.	Myrica fallax Lesquereux	Myrica fallax Lesquereux Myrica fallax (0.9) Myrica fallax DC. (0.81) Create new item
12.	Planera longifolia var. myricaefolia Lesquereux	Planera longifolia var. myricaefolia Lesquereux Planera longifolia myricaefolia (0.9) Planera longifolia subsp. myricaefolia (Lesquereux, 1883) (0.9) Create new item
13.	Yoldia gala Woodring	Yoldia gala Woodring and Bramlette 1950 Choose new match
14.	Schmidtella crassimarginata Ulrich	Schmidtella crassimarginata Ulrich 1890 Choose new match
15.	Rissoa (Onoba) marylandica Martin	Rissoa (Onoba) marylandica Martin 1904 Choose new match

Figure 2.

A basic reconciliation example where rows 11 - 12 require a human to make a choice, while rows 13 - 15 reconciled automatically.

Beyond the basic reconciliation (as seen in Fig. 2) there are more data that researchers are interested in. The Service Protocol allows one to add optional "extended" features. For example, for scientific names, we provide "currently accepted" names, data sources where a name was found, taxonomic classification, etc. Fig. 3. To make these fields standardized, we would need a recommendation document that describes additional fields and their format. We need interested parties to participate in its creation and agree on its usage. The same would apply to optional input filters, for example for restricting reconciliation to certain data sources or higher taxonomic entities.



Add columns from reconciled column Reconcile

Add property

Suggested properties

- AllDataSources
- CanonicalForm
- Classification
- CurrentName
- DataSource
- OutlinkURL

Preview Reset

Reconcile	CanonicalForm remove configure	DataSource remove configure
Myrica fallax	Myrica fallax	Tropicos
<not reconciled>		
Yoldia gala Woodring and Bramlette 1950	Yoldia gala	PaleoBioDB
Schmidtella crassimarginata Ulrich 1890	Schmidtella crassimarginata	PaleoBioDB
Rissoa (Onoba) marylandica Martin 1904	Rissoa marylandica	PaleoBioDB
Amalthea marylandica Martin, 1904	Amalthea marylandica	GBIF Backbone Taxonomy

Figure 3.

Extended reconciliation fields.

We think OpenRefine would be a significant step forward for standardization between name-reconciliation tools.

Keywords

name-reconciliation services, interoperability, FAIR, Global Names Verifier

Presenting author

Dmitry Mozzherin

Presented at

SPNHC-TDWG 2024

Acknowledgements

Special thanks to David Shorthouse and Nicky Nicolson for their encouragement, advice, and help in implementation of this OpenRefine name-reconciliation tool.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Islam S, Papale D, Vaira L, Rosati I, Peterseil J, Pichot C (2024) Navigating taxonomic complexity: A use-case report on FAIR scientific name-matching service usage in ENVRI Research Infrastructures. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e121871
- Whitmire A, Mozzherin D (2023) Reconciling taxonomic names in OpenRefine via Global Names. URL: https://github.com/gnames/gnverifier/wiki/OpenRefine-readme