Enhancing Herbarium Systems Using Name Matching Mechanisms

Walter G. Berendsohn[‡,§], Silvia Lusa Bernal[|], Banessa Falcón Hidalgo[¶], Dagoberto Rodríguez Delcid[#], Peter W. Moonlight[¤], Henry Riley Engledow[«]

‡ Freie Universitaet Berlin, Berlin, Germany
§ Berlin Botanic Garden, Berlin, Germany
| Real Jardín Botánico-CSIC, Madrid, Spain
¶ Jardín Botánico Nacional, Universidad de La Habana, Havanna, Cuba
# Jardín Botánico La Laguna, Antiguo Cuscatlán, El Salvador
¤ Trinity College Dublin, the University of Dublin, Dublin, Ireland
« Botanic Garden Meise, Meise, Belgium

Corresponding author: Walter G. Berendsohn (w.berendsohn@bgbm.org)

Abstract

Virtual aggregators of organism names and taxa play a normative role in consolidating the global biodiversity information infrastructure, serving as resources for researchers worldwide. These aggregators may serve as the glue that binds together local data, ranging from individual researchers' spreadsheets to large databases containing taxonomic checklists for countries or entire regions of the world. The European Union-funded TETTRIs project (Transforming European Taxonomy through Training, Research and Innovations) targets both local data holders and aggregators, aiming to motivate and enable local users to verify their data with the aggregators and, optionally, to link to the aggregators' services (Berendsohn 2023). It also aims to foster the development of aggregator-side services that streamline usage and facilitate the establishment of such linkages.

Here we focus specifically on botanical collection databases, i.e., the plant names contained in herbarium databases. By matching these names to botanical data aggregators like the World Flora Online (WFO) Plant List or the World Checklist of Vascular Plants (WCVP), curators can identify and correct obvious errors in their records. Where an exact match is obtained, the curators can check the aggregator's opinion about the nomenclatural validity, the taxonomic acceptance and the classification of their name, which may help e.g., in the processing of loan requests. Additionally, by reporting missing or incorrect names or commonly used orthographic variants, curators contribute to improving the overall quality of the infrastructure.

There are several services that may be used for name matching. We strongly suggest using the service offered by the WFO Plant List, because the dataset is becoming the most comprehensive global resource for the names of plants (excluding algae). It is

inclusive, i.e., it tries to cover all names and name-like designations that have been used in published taxonomic sources. It provides unique, resolvable and stable WFO name identifiers (Miller et al. 2023), and assigns new ones for names that have been corrected. It is supported by a broad spectrum of international botanical institutions (52, up to now). It closely cooperates with existing nomenclators such as Kew's WCVP, Missouri Botanical Garden's TROPICOS, and the International Plant Names Index (IPNI). Large parts are edited by TENs (Taxonomic Expert Networks, see Borsch et al. 2020) that may respond to input from users. It provides a unique, stable, resolvable WFO name identifier that can be integrated into local databases and used to follow up on changes made on the aggregator's side. The data is published under a CC0 rights waiver and the ID connects it to the World Flora Online content website.

We looked at specimen occurrence data supplied by herbaria to the Global Biodiversity Information Facility (GBIF). GBIF facilitated a dataset (GBIF.ORG 2024) with a count of specimens for all distinct scientific names in the supplied data for Plantae, plus the name they had matched it to in the GBIF backbone (a pragmatic assemblage of name and taxa from multiple botanical data sources). We filtered out algal groups by selecting only names from Anthocerotophyta, Bryophyta, Marchantiophyta, and Tracheophyta. With the 110,089,964 specimen records came 2,922,635 distinct names. GBIF is taking a number of measures to match names, so that in the end most of the specimens can be assigned to a taxon name in the GBIF backbone. However, we focused on exact literal matches of names, including author abbreviations. Only 681.211 (23.3%) of the uploaded names had an exact match in the GBIF taxonomic backbone, representing 40,401,731 occurrences (36.7% of all). So there is clearly room for improvement.

Some measures can be taken on the aggregator's side to improve the matching process. For example, aggregators should optionally provide exact matches for minor discrepancies in name strings that do not reflect true differences, such as spacing in abbreviated author citations or removing designations like 'spp.' or 'Indet.' to match only the name-citing portion. Other corrections have to be made in the herbarium database itself, e.g., the addition of name authors where these are missing. We found that author citations or the lack of these, represented the main issue of non-exact matching in the data uploaded to GBIF.

With local databases and aggregators improving their data and services, a subscription service using aggregator IDs as outlined in Berendsohn 2023 may become a realistic possibility in the near future. TETTRIs is assembling a wish list for aggregator services, which also includes the issues we have identified with the GBIF data submissions. We will further analyse these data and those from the herbaria of the authors' institutions to suggest priorities in data cleaning measures. We will consider how herbarium feedback can be leveraged to enhance aggregator datasets. We will also suggest measures for general collection management systems to facilitate name matching, starting with those used in our herbaria (JACQ, Specify).

Keywords

herbarium management system, JACQ, Specify, WFO, WFO Plant List, Catalogue of Life

Presenting author

Walter G. Berendsohn

Presented at

SPNHC-TDWG 2024

Funding program

Horizon Europe Programme Grant Agreement 101081903

Grant title

TETTRIs, Transforming European Taxonomy through Training, Research and Innovations

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Berendsohn WG (2023) Use Cases for Scientific Name Identifiers and Name Matching: Progress report from the TETTRIs project. Biodiversity Information Science and Standards 7 https://doi.org/10.3897/biss.7.109666
- Borsch T, Berendsohn WG, Dalcin E, Delmas M, Demissew S, Elliott A, Fritsch P, Fuchs A, Geltman D, Güner A, Haevermans T, Knapp S, le Roux MM, Loizeau P, Miller C, Miller J, Miller J, Palese R, Paton A, Parnell J, Pendry C, Qin H, Sosa V, Sosef M, von Raab-Straube E, Ranwashe F, Raz L, Salimov R, Smets E, Thiers B, Thomas W, Tulig M, Ulate W, Ung V, Watson M, Jackson PW, Zamora N (2020) World Flora Online: Placing taxonomists at the heart of a definitive and comprehensive global resource on the world's plants. TAXON 69 (6): 1311-1341. https://doi.org/10.1002/tax.12373
- GBIF.ORG (2024) (28 August 2024) GBIF Occurrence Download. Release date: 2024-8-28. URL: https://doi.org/10.15468/dl.6v7m4m

- Miller C, Berendsohn WG, Ulate W, Hyam R (2023) WFO-IDs: Unique Identifiers for All Known Plants Managed by the World Flora Online. Biodiversity Information Science and Standards 7 https://doi.org/10.3897/biss.7.111210