GBIF-Compliant Data Pipeline for the Management and Publication of a Global Taxonomic Reference List of Pests in Natural History Collections

Carla Novoa Sepúlveda[‡], Stephan Biebl[§], Nadja Pöllath[|], Stefan Seifert[¶], Markus Weiss[¶], Tanja Weibulat[¶], Dagmar Triebel[¶,‡]

‡ Staatliche Naturwissenschaftliche Sammlungen Bayerns, Botanische Staatssammlung München, Munich, Germany
§ Ingenieurbüro für Holzschutz, Benediktbeuern, Germany
| Staatliche Naturwissenschaftliche Sammlungen Bayerns, Staatssammlung für Paläoanatomie München, Munich, Germany
¶ Staatliche Naturwissenschaftliche Sammlungen Bayerns, SNSB IT Center, Munich, Germany

Abstract

There is a growing demand for monitoring pests in natural history collections (NHCs) and establishing integrated pest management (IPM) solutions (Crossman and Ryde 2022). In this context, up-to-date taxonomic reference lists and controlled vocabularies following standard schemes are crucial and facilitate recording organisms detected in collections.

The data pipeline described here results in the publication of a taxon reference list based on information from online resources and standard IPM literature. Most of the over 140 pest taxa on species level and above are insects, the rest belong to other animal groups and fungi.

The complete taxon names, synonyms, English and German common names, and the hierarchical classification (parent-child relationships) are organised in a client-server installation of DiversityTaxonNames (DTN) at the Bavarian Natural History Collections (SNSB). DTN is a Microsoft Structured Query Language (MS SQL) database tool of the Diversity Workbench (DWB) framework with a published Entity Relation (ER) diagram

(Hagedorn et al. 2019). The management is done using the Global Biodiversity Information Facility (GBIF) backbone taxonomy as external name resource, with linkage to the respective Wikidata Q item ID as a external persistent identifier (PID). Moreover, information on pest occurrence in NHCs is given, distinguishing the Consortium of European Taxonomic Facilities (CETAF) major NHC collection types affected (i.e., heritage sciences, life sciences and earth sciences) and the object categories, e.g., natural objects/ specimens damaged. The data management in DTN enables the long-running curation, done by list curators.

The generic data pipeline for the management and publication of a Global Taxonomic Reference List of Pests in NHCs is based on the DTN taxon lists concept and architecture and described under About "Taxon list of pest organisms for IPM at natural history collections compiled at the SNSB". It includes four steps (A–D) with significant results for best practices of data processing (Fig. 1).
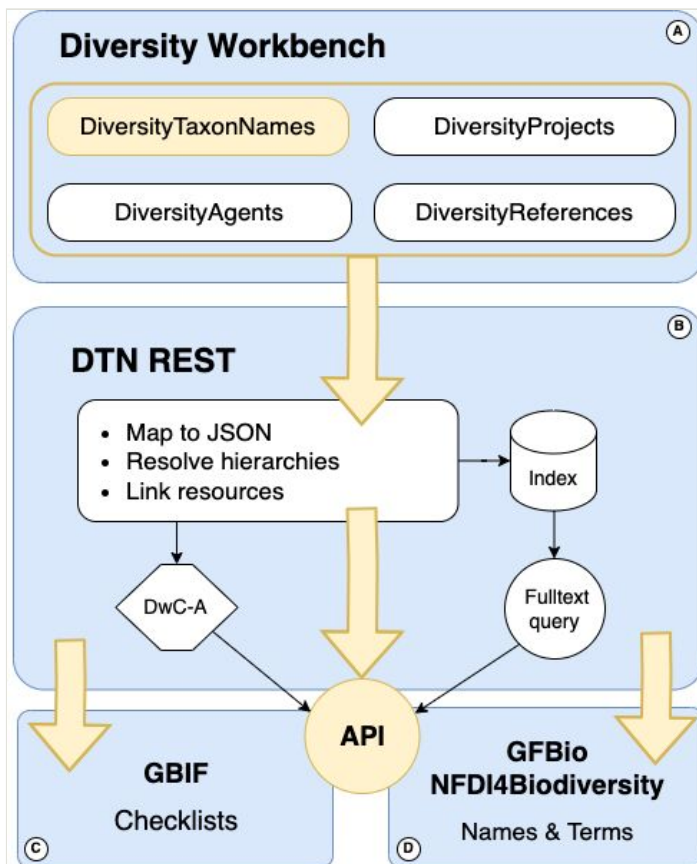

Figure 1.

Generic data management and publication in four steps (A-D) as applied for the IPM taxon reference list (Novotný et al. 2022)

A. The data is managed and processed for publication by list curators in the database DiversityTaxonNames (DTN).

As a result, the list can be kept up-to-date and is—without transformation—ready to be used for IPM solutions at any NHC with a DiversityCollection installation and as part of the DWB cloud services.

B. The up-to-date data is publicly available via the DTN REST Webservice for Taxon Lists with machine-readable Application Programming Interface (API).

As a result, the dynamic list publication service can be used as a reference backbone for establishing IPM solutions for pest monitoring at any NHC.

C. The data is provided via the GBIF checklist data publication pipeline of the SNSB through GBIF validation tools and Darwin Core Archive in DwC-A (zip format) for GBIF.

As a result, the checklist information becomes part of the GBIF network with GBIF ChecklistBank and GBIF Global Taxonomy. This ensures future compliance of data with the Findability, Accessibility, Interoperability, and Reuse (FAIR) guiding principles.

D. The DTN REST Web service for Taxon Lists (currently 60 lists) is registered and accessible through the German Federation for Biological Data (GFBio) Terminology service.

As a result, the lists with external PIDs and other information are available as a service (see DTN lists overview). In the upcoming Research Data Commons of the German National Research Data Infrastructure (NFDI) Initiative (Diepenbroek et al. 2021), it will be part of a standardized layer of APIs with an agreed interface scheme for improved accessibility.

The provided tools, API and data are part of the upcoming NFDI4Biodiversity service portfolio. Future scenarios include the use of the list items and properties as classes for diagnosis purposes with DiversityNaviKey (Triebel et al. 2021) including the publication of images for identifying pests.

Keywords

integrated pest management, Diversity Workbench, taxon names, web service, API, research data infrastructure, name identifiers, NFDI

Presenting author

Carla Novoa Sepúlveda

Presented at

Funding program

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Crossman A, Ryde S (Eds) (2022) Integrated Pest Management for Collections, Proceedings of the 2021: A Pest Odyssey, the Next Generation. 264 pp. [In English]. [ISBN 9781909492837].
- Diepenbroek M, Schimmler S, Ebert B (2021) Sektionskonzept Common Infrastructures zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V. Zenodo https://doi.org/10.5281/zenodo.5607489
- Hagedorn G, Weiss M, Triebel D (2019) DiversityTaxonNames information model (version 2.5.29). https://diversityworkbench.net/Portal/DiversityTaxonNamesModel_2.5.29. Accessed on: 2023-7-25.
- Novotný P, Seifert S, Rohn M, Diewald W, Štech M, Triebel D (2022) Software infrastructure and data pipelines established for technical interoperability within a cross-border cooperation for the flora of the Bohemian Forest. Biodiversity Data Journal 10 https://doi.org/10.3897/bdj.10.e87254
- Triebel D, Grunz A, Seifert S, Link A, Rambold G (2021) DiversityNaviKey, a Progressive Web Application for interactive diagnosis and identification. INFORMATIK 2021 - Computer Science & Sustainability. 517-538. https://doi.org/10.18420/informatik2021-040