A Possible Workflow from New and Legacy Publications to keep the World Flora Online up to date with New Species and Augmenting Taxonomic Treatments

Donat Agosti[‡], Laurence Benichou[§], Lyubomir Penev[|,¶], Roger Hyam[#]

‡ Plazi, Bern, Switzerland
§ National Museum of Natural History, Paris, France
| Pensoft Publishers, Sofia, Bulgaria
¶ Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences, Sofia, Bulgaria
# Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom

Corresponding author: Donat Agosti (agosti@plazi.org)

Abstract

Thousands of new species are discovered each year, and new results are published to add to the knowledge of existing species. A growing number of these are immediately accessible through the Biodiversity Literature Repository (BLR) and reused by Global Biodiversity Information Facility (GBIF), bringing the number of treatments covering plant species to over 25,000 treatments. This includes the findable, accessible, interoperable, and resuable (FAIR) treatments and related figures, and in many cases the material citation of the holotype, and links to the collection, specimen and gene sequences attributed to the codes. The FAIR data is deposited in the Biodiversity Literature Repository ensuring long-term access, and includes rich, customized metadata describing its content using standard vocabularies (e.g. Darwin Core (DwC) or Open Biological and Biomedical Ontology (OBO) Foundry, as well as links to related items and data reuse (e.g. GBIF and CheckListbank).

TreatmentBank is a European Research Infrastructure, which annotates taxonomic publications including quality control, as much as possible by machine, and as needed by human curation. The process is versatile, the output can be customized and immediately used by GBIF and the Swiss Institute of Bioinformatics library system (SIBiLS) to add treatments akin to articles in PubMed Central. Another focus of conversion could be floras (several series published in Muséum national d'Histoire naturelle including those that are currently only available in print). In addition to the articles provided through these collaborations, other journals were processed, totaling 5,900 articles with 54,000 treatments of which 5,600 are plants, 48,000 figures, 140,000 material citations, of which 9,000 (670 plants) are new species and 670 (16) new genera described in 2021. Arcadia F

und is funding a project to include over 100 additional journals into the daily processing, as well as annotating data from legacy publications, with the possibility of including more journals covering plant taxonomy. The workflow is based on the open source software GoldenGate Imagine, a tool used to convert articles from PDF format into an Image Markup File (IMF) export file including all the annotations in a set of CSV files and allowing checking and improve the quality of data. The quality control parameter can be defined according to specific user requirements and the curation of the annotation can be done as part of the workflow or by third parties. The IMF can be uploaded to TreatmentBank where FAIR data is generated through their deposits in the Biodiversity Literature Repository. All the data is accessible from the Plazi API in various formats such as XML, DwC-A (DarwinCore Archive) or RDF, and checked when the data has been deposited or reused by returning the link of the respective recipient's reuse identifier. Furthermore, the treatment citations annotated in the publications provide access to the treatments related to the same taxon, including synonyms.

In this presentation, we explain the data annotated and made FAIR by TreatmentBank in order to discuss what would be needed to import this rapidly increasing corpus of taxonomic treatments as a new source for WFO (World Flora Online). A side effect of the discussion could be suggestions to publishers on how they should provide data so that the taxonomic data can be automatically integrated into WFO to keep it as up-to-date as possible.

Keywords

data import, publishing, collaboration, literature, biodiversity

Presenting author

Donat Agosti

Presented at

TDWG 2022

Acknowledgements

TreatmentBank collaborates with Pensoft publishers, the Muséum national d'Histoire naturelle, Paris (MNHN) and the CETAF e-publishing working group's journal European Journal of Taxonomy.

Funding program

The TreatmentBank infrastructure is supported by the Horizon Europe funded project Biodiversity Community Integrated Knowledge Library (BiCIKL), the Arcadia Fund and the Swissuniversities funded eBioDiv project.

Conflicts of interest