Biodiversity Information Science and Standards 4: e59179
doi: 10.3897/biss.4.59179

Conference Abstract

Liberating the Richness of Facts implicit in taxonomic Publication: The Plazi Workflow

Donat Agosti[‡], Marcus Guidoti[§], Guido Sautter[|]

[‡] Plazi, Bern, Switzerland
[§] Plazi, Porto Alegre, Brazil
[|] IPD Böhm, Karlsruhe Institute of Technology, Karlsruhe, Germany

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

Received: 30 Sep 2020 | Published: 09 Oct 2020

Citation: Agosti D, Guidoti M, Sautter G (2020) Liberating the Richness of Facts implicit in taxonomic Publication: The Plazi Workflow . Biodiversity Information Science and Standards 4: e59179. https://doi.org/10.3897/biss.4.59179

Abstract

The growing corpus of hundreds of millions of pages of taxonomic literature reporting research results based on specimens is very rich in facts. In order to make them reusable, Plazi, Pensoft and Zenodo are building and maintaining the Biodiversity Literature Repository which includes a workflow to discover, describe, store, in order to making these facts open access, findable, accesible, interoperable and reusable (FAIR).

Currently, 43,000 articles have 406,000 material citations, and around 50% of annually new described species are made accessible and immediately reused by the Global Biodiversity Information Facility (GBIF). All the images are deposited at the Biodiversity Literature Repository (BLR), as well as the taxonomic treatments. For each of these deposits enriched metadata is added and a Digital Object Identifier (DOI) is minted. Through this process, Plazi is the single largest data set provider to GBIF and continues to provide ca. 45,000 unique taxonomic names at GBIF.

The workflow is optimized for born digital portable data format (PDF) based publications, but other formats can also be ingested, including TaxPub, a taxonomy specific version of the Journal Article Tag Suit (JATS) XML. After ingestion, the PDF is readily converted to an open-access, proprietary format called Image Markup File (IMF). IMF is a compressed file format that consists of the enhanced information contained in the PDF, with figures and

tables properly extracted. The IMFs are then housed at TreatmentBank, with associated exported files, including DwC-A for each parent article and their respective taxonomic treatments, XMLs of treatments and GBIF datasets of their parent articles. Taxonomic treatments, in addition to figures and the original PDFs, are also deposited on Zenodo, where a DOI is minted if none is already available. These Zenodo deposits include in the metadata links back to the different data and file formats, including the treatments XMLs, maintaining the system connected and up-to-date. Third-party players, like GBIF, Global Biotic Interactions (GloBI), Ocellus, OpenBiodiv and Synospecies are constantly fed by system hookups, which guarantees data consistency after further edits.

The PDF-IMF conversion and data enhancement is possible due to Plazi's open-source software called GoldenGate Imagine. Ingested XMLs, that are validated against the TaxPub scheme, follow a similar path into the system and the many third-party applications.

This operation is supported by the Arcadia Fund as well by service contracts from publishers to disseminate their data. In addition, the workflow has been contributing treatments, images from numerous publications relevant to understanding the virus spillover as part of the CETAF COVID19 task force.

In this lecture this workflow is described and explained, including the associated infrastructure and its ongoing changes and upcoming steps of development.

Keywords

biodiversity, digital library, FAIR data, open access

Presenting author

Donat Agosti

Presented at

TDWG 2020