Pre-Publication Data Linking in Taxonomy and Biodiversity: The ARPHA and Metotaxa-Metostem Publishing Systems

Laurence Benichou[‡,§], Marianne Salaün[‡], Iva Boyadzhieva[|], Seyhan Demirov[|], Teodor Georgiev[|], Lyubomir Penev[¶,#]

‡ Museum national d'Histoire naturelle, Paris, France
§ CETAF, Brussels, Belgium
| Pensoft Publishers, Sofia, Bulgaria
¶ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
# Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria

Corresponding author: Laurence Benichou (laurence.benichou@mnhn.fr)

Abstract

The traditional way of publishing in PDF makes it difficult to retrospectively convert the legacy literature into data. This presentation will discuss pre-publication tagging as an alternative solution for publishing FAIR (Findable, Accessible, Interoperable, Resuable) biodiversity data.

The Metotaxa-Metostem workflow

The MetoTaxa project aims to create a new digital production chain for the *European Journal of Taxonomy*, which enables the pre-publication semantic structuring of text, automatic tagging and semantic enrichment (annotation).

The system is based on a single-source publishing model, where the development of an XML file enables technical editors to automatically enrich text and produce multiple digital outputs. This makes it possible to structure generic or domain-specific sections of articles (e.g., Introduction; Material and methods; Taxon names or Material examined). Thanks to the GoldenGate API developed by Plazi, the Text Encoding Intiative (TEI) XML source file is automatically annotated with JATS TaxPub tags: taxon names are labeled and each authorship can be checked via Catalogue of Life, each element of the material examined is parsed thanks to the preformatting of the text (Chester et al. 2019). Also, each bibliographic reference is parsed into Journal Article Tag Suite (JATS) elements (author names, title, journal, etc.), which automatically links references to their in-text citations. Pre-publication tagging will be carried out by the technical editors and then checked by the authors before publication, and will be sent to databases such as Global Biodiversity Information Facility (GBIF) or Biodiversity Literature Repository (BLR) as soon as the article is published.

We will also briefly present MetoStem, which offers a technical solution for the digital transformation of monographs, and particularly floras. The tools and methods developed by this project will enable advanced publication of interoperable structured text and data.

ARPHA Publishing Platform

Launched in 2010 by Pensoft, ARPHA (Penev et al. 2010) is the first ever scholarly publishing platform to support pre-publication semantic tags and enhancements to entities (e.g., taxon treatments, taxon names, sequences) in the JATS TaxPub XML format developed by Plazi, which are then embedded into the HTML version of the article. Having proved advantageous for biodiversity scientists, Pensoft's pre-publication tagging workflow has since been adopted by over 30 biodiversity journals hosted on ARPHA.

The second development stage of ARPHA was marked by the launch of ARPHA Writing Tool (AWT)*[1] and Biodiversity Data Journal in 2013. AWT supports import of Darwin Core structured data from GBIF, Barcode of Life Data Systems (BOLD) and Integrated Digitized Biocollections (iDigBio) directly into manuscripts. These are also exported automatically as published material citations to GBIF. AWT also provides several other unique tools encompassed within the ARPHA-BioDiv toolbox (Penev et al. 2017).

Currently, AWT is being redeveloped into a standalone, freely accessible installation (AWT 2.0), based on a micro-service architecture. It enables new semantic enhancements during the authoring process, which can be confirmed by the authors before manuscript submission. Such enhancements include the in-text citations context by CiTO ontology; automated tagging of taxon names and linking to their identifiers in authoritative sources; annotator tool; nanopublication module; automated search and import of references; treatment citation module; export/import to/from JATS TaxPub; and internal communication tool for contributors.

Keywords

semantic publishing, data dissemination, data liberation

Presenting author

Laurence Bénichou

Presented at

TDWG 2023

Funding program

Part of the work was supported by the Biodiversity Community Integrated Knowledge Library (BiCIKL) project, funded by the European Union's Horizon 2020 under grant agreement No 101007492; MetoStem project funded by the Fond National pour la Science ouverte No.2, France.

Grant title

BiCIKL - Biodiversity Community Integrated Knowledge Library

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Chester C, Agosti D, Sautter G, Catapano T, Martens K, Gérard I, Bénichou L (2019) EJT editorial standard for the semantic enhancement of specimen data in taxonomy literature. European Journal of Taxonomy 586 https://doi.org/10.5852/ejt.2019.586
- Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, Smith V, Brake I, Ryrcroft S, Scott B, Johnson N, Morris R, Sautter G, Chavan V, Robertson T, Remsen D, Stoev P, Parr C, Knapp S, Kress WJ, Thompson F, Erwin T (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1-16. https://doi.org/10.3897/zookeys.50.538
- Penev L, Georgiev T, Geshev P, Demirov S, Senderov V, Kuzmova I, Kostadinova I, Peneva S, Stoev P (2017) ARPHA-BioDiv: A toolbox for scholarly publication and dissemination of biodiversity data based on the ARPHA Publishing Platform. Research Ideas and Outcomes 3 https://doi.org/10.3897/rio.3.e13088

Endnotes

*1 https://arpha.pensoft.net/