Post-Publication Linking

Felipe Lorenz Simoes[‡], Donat Agosti[§], Diego Janisch Alvares[‡], Jonas Blanco Castro[‡], Julia Giora[‡], Valdenar da Rosa Gonçalves[‡], Tatiana Petersen Ruschel[‡], Carolina Sokolowicz[‡], Juliana Mariani Wingert[‡]

‡ Plazi, Porto Alegre, Brazil
§ Plazi, Bern, Switzerland

Corresponding author: Felipe Lorenz Simoes (simoes@plazi.org)

Abstract

One of the main challenges in biodiversity data reusability is finding ways to transform what is provided in research publications into different and reusable formats, following the FAIR (Findable, Accessible, Interoperable, Reusable) principles (Agosti and Egloff 2009).

Most often, data is restricted to text, figures and tables in the so-called “PDF prison” or other flat formats. Plazi's infrastructure and workflow (Guidoti et al. 2021) transform such data into reusable formats that can then be exported and linked across different platforms, such as the Global Biodiversity Information Facility (GBIF), Biodiversity Literature Repository, Zenodo, Synospecies, ChecklistBank, and OpenBiodiv among others.

In order to liberate the many relevant pieces of information, such as taxonomic treatments (Catapano 2019), material citations (Darwin Core term MaterialCitation) or bibliographic references from the publication types mentioned above, one has to run a single document or a batch of documents through a series of extraction steps, which can be done manually or automatically, through the use of templates. The latter are a set of parameters that tell the Plazi-dedicated software (GoldenGATE suite) how to read and where to find key pieces of information; these parameters are established by examining publication standards and publisher-specific layouts, followed by a series of iterative tests, to ascertain the quality of the automation.

However, even with a high number of tests to ensure a better extraction, human quality control is still needed (Simoes et al. 2021). To that end, Plazi has a quality control process, based on logical rules, which checks the components of the extracted document, flagging errors in four different levels of severity, which can then be checked and corrected (if needed) by a trained user. These errors are also used in a data transit control mechanism, internally dubbed “the gatekeeper”, which blocks certain data transits to create deposits or reuse of data in the presence of specific errors.

In this presentation, we will go through the steps of the entire process, from publication to liberated data (and how it is presented in the linked platforms), highlighting the importance of accurate quality control, and explore some of the many challenges along the way.

Keywords

annotations, biodiversity, data-oriented, process, workflow, strategy, digital library

Presenting author

Felipe Lorenz Simões

Presented at

TDWG 2023

Funding program

The Biodiversity Community Integrated Knowledge Library (BiCIKL) project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492. Plazi acknowledges the support from the Arcadia Fund.

Grant title

Biodiversity Community Integrated Knowledge Library (BiCIKL)

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Agosti D, Egloff W (2009) Taxonomic information exchange and copyright: the Plazi approach. BMC Research Notes 2 (1). https://doi.org/10.1186/1756-0500-2-53
- Catapano T (2019) TaxPub: An Extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions. Journal Article Tag Suite Conference (JATS-Con), Bethesda (MD), USA. https://doi.org/10.5281/zenodo.3484285
- Guidoti M, Sokolowicz C, Simoes F, Gonçalves V, Ruschel T, Alvares DJ, Agosti D (2021) TreatmentBank: Plazi's strategies and its implementation to most efficiently

liberate data from scholarly publications. Biodiversity Information Science and Standards 5: 75690. https://doi.org/10.3897/biss.5.75690

- Simoes F, Agosti D, Guidoti M (2021) Delivering Fit-for-Use Data: Quality control. Biodiversity Information Science and Standards 5: 75432. https://doi.org/10.3897/biss.5.75432