Delivering Fit-for-Use Data: Quality control

Felipe Simoes[‡], Donat Agosti[§], Marcus Guidoti[‡]

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

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

Abstract

Automatic data mining is not an easy task and its success in the biodiversity world is deeply tied to the standardization and consistency of scientific journals' layout structure. The various formatting styles found in the over 500 million pages of published biodiversity information (Kalfatovich 2010), pose a remarkable challenge towards the goal of automating the liberation of data currently trapped on the printed page. Regular expressions and other pattern-recognition strategies invariably fail to cope with this diverse landscape of academic publishing. Challenges such as incomplete data and taxonomic uncertainty add several additional layers of complexity.

However, in the era of big data, the liberation of all the different facts contained in biodiversity literature is of crucial importance. Plazi tackles this daunting task by providing workflows and technology to automatically process biodiversity publications and annotate the information therein, all within the principles of FAIR (findable, accessible, interoperable, and reusable) data usage (Agosti and Egloff 2009). It uses the concept of taxonomic treatments (Catapano 2019) as the most fundamental unit in biodiversity literature, to provide a framework that reflects the reality of taxonomic data for linking the different pieces of information contained in these taxonomic treatments. Treatment citations, composed of a taxonomic name and a bibliographic reference, and material citations carrying all specimen-related information are additional conceptual cornerstones for this framework. The resulting enhanced data are added to TreatmentBank. Figures and treatments are made Findable, Accessible, Interoperable and Reuseable (FAIR) by depositing them including specific metadata to the Biodiversity Literature Repository community (BLR) at the European Organization for Nuclear Research (CERN) repository Zenodo, and pushed to GBIF. The automation, however, is error prone due to the constraints explained above.

In order to cope with this remarkable task without compromising data quality, Plazi has established a quality control process, based on logical rules that check the components of the extracted document raising errors in four different levels of severity. These errors are also used in a data transit control mechanism, “the gatekeeper”, which blocks certain data transits to create deposits (e.g., BLR) or reuse of data (e.g., GBIF) in the presence of specific errors. Finally, a set of automatic notifications were included in the plazi/

community Github repository, in order to provide a channel that empowers external users to report data issues directly to a dedicated team of data miners, which will in turn and in a timely manner, fix these issues, improving data quality on demand.

In this talk, we aim to explain Plazi's internal quality control process and phases, the data transits that are potentially affected, as well as statistics on the most common issues raised by this automated endeavor and how we use the generated data to continuously improve this important step in Plazi's workflow.

Keywords

annotations, biodiversity data, FAIRification, TreatmentBank

Presenting author

Felipe Simoes

Presented at

TDWG 2021

Funding program

The BiCIKL (Biodiversity Community Integrated Knowledge Library — https://bicikl-project.eu/) project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492

Conflicts of interest

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. Zenodo https://doi.org/10.5281/zenodo.3484285
- Kalfatovich M (2010) BHL Australia Kick Off Meeting: Melbourne Museum. 1 June 2010. Melbourne, Australia. URL: https://www.slideshare.net/Kalfatovic/3-years-on-the-biodiversity-heritage-library?qid=3a0bdbbc-8b89-4260-a69d-93b58c8c6885&v=&b=&from_search=19