BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37046
doi: 10.3897/biss.3.37046

OPEN ACCESS

Conference Abstract

The Plazi Workflow: The PDF prison break for biodiversity data

Donat Agosti[‡], Terry Catapano[§], Guido Sautter[§], Willi Egloff[§]

‡ www.plazi.org, Bern, Switzerland
§ Plazi, Bern, Switzerland

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

Received: 10 Jun 2019 | Published: 13 Jun 2019

Citation: Agosti D, Catapano T, Sautter G, Egloff W (2019) The Plazi Workflow: The PDF prison break for biodiversity data. Biodiversity Information Science and Standards 3: e37046. https://doi.org/10.3897/biss.3.37046

Abstract

The Swiss NGO Plazi (http://plazi.org) has developed an automated workflow for liberating data, including images and text, from new taxonomic publications issued in PDF format. This stepwise process extracts, article metadata, illustrations and their captions, bibliographic references, scientific names, named geographic entities such as coordinates and country names, collection codes, and finally, taxonomic treatments. These extracted data are enhanced and published in TreatmentBank (http://plazi.org) and deposited in Biodiversity Literature Repository (https:/biolitrepo.org) respectively, in which a Digital Object Identifier (DataCite DOI) is minted for articles as well as their contained figures and taxon treatments, each linked to each other in their metadata. This input is complemented by the import of Journal Article Tag Suite/Taxpub XML based publications from Pensoft publishers (e.g. Zookeys, Journal of Hymenoptera Research; https://pensoft.net/browse_journals) that are semantically enhanced during their journal production workflow. Upon import, materials citation are discovered and parsed, and the taxonomic treatments added to TreatmentBank where a persistent identifier is minted. From TreatmentBank data from taxonomic treatments, including occurence data from cited specimens, are submitted to GBIF (http://gbif.org), or are accessible via API. Treatments and material citations from more than 26,200 articles have been registered. The articles can be found on GBIF using the Digital Object Identifier in the search field.

Plazi, together with Pensoft Publishers, has processed over 26,000 articles containing more than 284,000 taxonomic treatments, 190,000 images, 50,000 georeferenced materials citations, together comprising an estimated 100 million facts. Through the support of the Arcadia Fund (https://www.arcadiafund.org.uk/) Plazi's processing is expanding to cover a sufficient number of journals to liberate the data of over 50% of the new described animal species annually. This will complement an existing service provided to the Muséum National d'Histoire Naturelle, Paris, to convert the European Journal of Taxonomy and their other journals (http://sciencepress.mnhn.fr/en/periodiques/adansonia/40/1) to JATS/TaxPub (https://www.ncbi.nlm.nih.gov/books/NBK47081), as well as an increasing portfolio of journals published in JATS/TaxPub by Pensoft Ltd.

Keywords

publishing, taxonomic treatments, data conversion, taxonomy, TaxPub

Presenting author

Donat Agosti

Presented at

Biodiversity_Next 2019