Synospecies, a Linked Data Application to Explore Taxonomic Names

Reto Gmür‡, Donat Agosti‡, Guido Sautter‡

‡ Plazi, Bern, Switzerland

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

Abstract

Synospecies is a linked data application to explore changes in taxonomic names (Gmür and Agosti 2021). The underlying source of truth for the establishment of taxa, the assignment and re-assignment of names, are taxonomic treatments. Taxonomic treatments are sections of publications documenting the features or distribution of taxa in ways adhering to highly formalized conventions, and published in scientific journals, which shape our understanding of global biodiversity (Catapano 2010). Plazi, a not-for-profit organization dedicated to liberating knowledge, extracts the relevant information from these treatments and makes it publicly available in digital form. Depending on the original form of a publication, a treatment undergoes several steps during its processing. All these steps affect the available digital artifacts extracted from the treatment's original publication. The treatments are digitalized, the text is annotated with a specialized editor, and cross-referenced and enhanced with other sources (Agosti and Sautter 2018). After these steps, the annotated text is transformed to the different structured data-formats used by other digital biodiversity platforms (e.g., Global Biodiversity Information Facility: Plazi.org taxonomic treatment database using Darwin Core Archive, generic linked data tools (e.g. lod view; RDF2h Browser) and other consuming applications (e.g Ocellus via Zenodeo using XML; openBioDiv using XML; HMW using XML; Biotic interaction browser using TaxPub XML; opendata.swiss using RDF) .

While these transformations have been taking place for a long time now, Plazi is now experimenting with making this process more transparent: with the Plazi Actionable Accessible Archive (PAAA) architecture both addition and modification of the digitalized treatments trigger an extensible set of workflows that are immediately executed on the GitHub platform. Not only is the exact definition and code of every workflow publicly accessible, but the results, errors and execution time of every single workflow is accessible as well. This offers an unprecedented degree of transparency and flexibility in the data processing that we have prototypically implemented for the creation of the RDF data used by Synospecies. As with the W3C GRDDL recommendation (https://www.w3.org/TR/grddl/) XSLT is used to transform XML to RDF/XML, a concrete syntax of the early days of RDF still supported by most RDF tools, allowing the data to be read as RDF. The used

XSLT document is part of the bundled gg2rdf GitHub action (https://github.com/plazi/gg2rdf) together with the other transformation steps required to generate a transformation result in the both human- and machine-readable RDF Turtle format. On the GitHub Actions page of the treatments-xml repository (https://github.com/plazi/treatments-xml/actions) one can see that every commit to this repository triggers a workflow run that takes approximately 12 minutes to execute. After that the transformation results are available in the treatments-rdf repository (https://github.com/plazi/treatments-rdf/). The commit of RDF data to the treatments-rdf repository triggers a webhook that loads the newly added data to the Plazi triplestore making it virtually immediately available in Synospecies.

Keywords

biodiversity, RDF, knowledge graph, treatment, citation, ontology, SPARQL, TreatmentBank

Presenting author

Reto Gmür

Presented at

TDWG 2022

Funding program

The TreatmentBank infrastructure is supported by the Horizon Europe funded project Biodiversity Community Integrated Knowledge Library (BiCIKL), the Arcadia Fund and the Swissuniversities funded eBioDiv project.

BiCIKL https://bicikl-project.eu/ grant number 101007492

Arcadia Fund https://www.arcadiafund.org.uk/

Swissuniversities https://www.swissuniversities.ch/themen/digitalisierung/open-science-2021-2024

eBioDiv https://ebiodiv.org/

Conflicts of interest

References

- Agosti D, Sautter G (2018) Text And Data Mininingworkflow To Make Scientific Publications Accessible. Zenodo https://doi.org/10.5281/zenodo.1288280
- Catapano T (2010) TaxPub: An Extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions. Zenodo https://doi.org/10.5281/zenodo.3484285
- Gmür R, Agosti D (2021) Synospecies, an application to reflect changes in taxonomic names based on a triple store based on taxonomic data liberated from publication . Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.75641