The BHL-Plazi Partnership: Getting data from the 1800s directly into 21st century, reused digital accessible knowledge

Diego Janisch Alvares[‡], Marcus Guidoti[‡], Felipe Simoes[‡], Carolina Sokolowicz[‡], Donat Agosti[§]

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

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

Abstract

Plazi is a Swiss non-governmental organization dedicated to the liberation of data imprisoned in flat, dead-end formats such as PDFs. In the process, the data therein is annotated and exported in various formats, following field-specific standards, facilitating free access and reutilization by several other service providers and end-users. This data mining and enhancement process allows for the rediscovery of the known biodiversity since the knowledge on known taxa is published into an ever-growing corpus of papers, chapters and books, inaccessible to the state-of-the-art service providers, such as Global Biodiversity Information Facility (GBIF). The data liberated by Plazi focuses on taxonomic treatments, which carry the unit of knowledge about a taxon concept in a given publication and can be considered the building block of taxonomic science. Although these extracted taxonomic treatments can be found in Plazi's TreatmentBank and Biodiversity Literature Repository (BLR), hosted in the European Organization for Nuclear Research (CERN) digital repository Zenodo, data included in treatments (e.g., material citations and treatment citations) can also be found in other applications as well, such as Plazi's Synospecies, Zenodeo, and GBIF. Plazi's efforts result in more Findable, Accessible, Interoperable, and Reusable (FAIR) biodiversity literature, improving, enhancing and enabling access to data included therein as digital accessible data, otherwise almost unreachable.

The Biodiversity Heritage Library (BHL), on the other hand, provides a pivotal service by digitizing heritage literature and current literature for which BHL negotiates permission, and provides free access to otherwise inaccessible sources.

In 2021, BHL and Plazi signed a Statement of Collaboration, aiming to combine the efforts of both institutions to contribute even further to FAIR-ifying biodiversity literature and data. In a collaborative demonstration project, we selected the earliest volumes and issues of the *Revue Suisse de Zoologie* in order to conduct a pilot study that combines the efforts of both BHL and Plazi.

The corpus is composed of eight volumes (tomes), 24 issues (numbers) and 98 papers, including a total of over 5000 pages and 200 images. To process this material, BHL assigned CrossRef Digital Object Identifiers (DOI) to these already digitally accessible publications. Plazi created a template to be used in GoldenGate-Imagine, indicating key parameters used for tailored data mining of these articles, and customized to the journal's graphic layout characteristics at that time. Then, we proceeded with quality control steps to provide fit-for-use data for BLR and GBIF by ensuring that the data was correctly annotated and eliminating potential data transit blockages at Plazi's built-in data gatekeeper. The data was then subsequently reused by GBIF. Finally, we present here the summary of the obtained results, highlighting the number of key publication attributes aforementioned (pages, images), but also including a drill-down into the different taxonomic groups, countries and collections of origin of the studied material, and more. All the data is available via the Plazi statistics, the Biodiversity Literature Repository Website and community at Zenodo, the Zenodeo APIs and GBIF where the data is being reused.

Keywords

biodiversity, digitally accessible knowledge, digital library, FAIR, Biodiversity Heritage Library, Biodiversity Literature Repository, GBIF

Presenting author

Donat Agosti

Presented at

TDWG 2021

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

Conflicts of interest