TreatmentBank: Plazi's strategies and its implementation to most efficiently liberate data from scholarly publications

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

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

Corresponding author: Marcus Guidoti (marcus.guidoti@gmail.com)

Abstract

Plazi's TreatmentBank is a research infrastructure and partner of the recent European Union-funded Biodiversity Community Integrated Knowledge Library (BiCIKL) project to provide a single knowledge portal to open, interlinked and machine-readable, findable, accessible, interoperable and reusable (FAIR) data. Plazi is liberating published biodiversity data that is trapped in so-called flat formats, such as portable document format (PDF), to increase its FAIRness. This can pose a variety of challenges for both data mining and curation of the extracted data. The automation of such a complex process requires internal organization and a well established workflow of specific steps (e.g., decoding of the PDF, extraction of data) to handle the challenges that the immense variety of graphic layouts existing in the biodiversity publishing landscape can impose. These challenges may vary according to the origin of the document: scanned documents that were not initially digital, need optical character recognition in order to be processed.

Processing a document can either be an individual, one-time-only process, or a batch process, in which a template for a specific document type must be produced. Templates consist of a set of parameters that tell Plazi-dedicated software how to read and where to find key pieces of information for the extraction process, such as the related metadata. These parameters aim to improve the outcome of the data extraction process, and lead to more consistent results than manual extraction. In order to produce such templates, a set of tests and accompanying statistics are evaluated, and these same statistics are constantly checked against ongoing processing tasks in order to assess the template performance in a continuous manner.

In addition to these steps that are intrinsically associated with the automated process, different granularity levels (e.g., low granularity level might consist of a treatment and its subsections versus a high granularity level that includes material citations down to named entities such as collection codes, collector, collecting date) were defined to accommodate

specific needs for particular projects and user requirements. The higher the granularity level, the more thoroughly checked the resulting data is expected to be.

Additionally, steps related to the quality control (qc), such as the “pre-qc”, “qc” and “extended qc” were designed and implemented to ensure data quality and enhanced data accuracy.

Data on all these different stages of the processing workflow are constantly being collected and assessed in order to improve these very same stages, aiming for a more reliable and efficient operation. This is also associated with a current Data Architecture plan to move this data assessment to a cloud provider to promote real-time assessment and constant analyses of template performance and processing stages as a whole.

In this talk, the steps of this entire process are explained in detail, highlighting how data are being used to improve these steps towards a more efficient, accurate, and less costly operation.

Keywords

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

Presenting author

Marcus Guidoti

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. Plazi acknowledges the support from the Arcadia Fund.

Conflicts of interest