Golden Gate Imagine – Text and data mining of taxonomic publications for taxonomic treatments and observation records

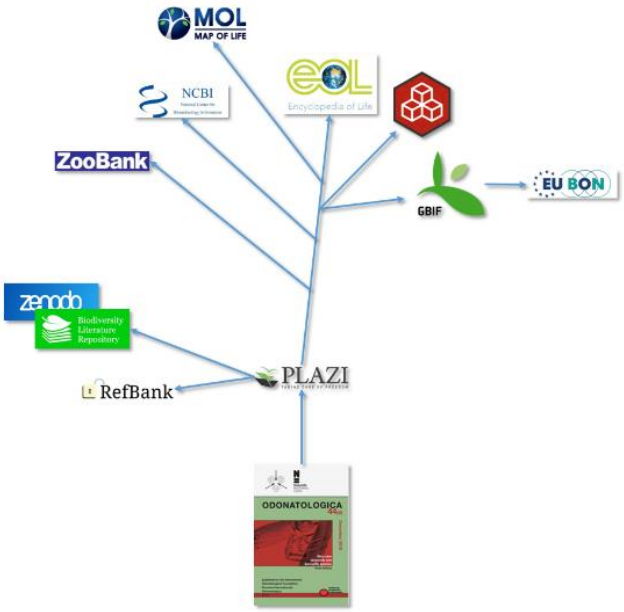
DATA MANAGEMENT

Donat Agosti (agosti@plazi.org)

Overview

Taxonomic literature contains not only the descriptions of biodiversity as we know it, but also citations of the specimens that form the basis for this primary taxonomic study. Because of their contribution to revisions, monographs, descriptions, and other primary taxonomic literature, these specimen records are among the highest quality biodiversity data available. In an age when we have the tools to effectively manage and analyse large quantities of data, and when environmental changes call for data-driven decision-making, the digitization, structuring and extraction of content from taxonomic literature is needed to provide a more comprehensive supply of information—especially for the many species not well covered in biodiversity data infrastructure. **GoldenGate Imagine (GCI)** allows semiautomatic, interactive extraction of taxonomic treatments, scientific names, named entities, bibliographic references, and observation data from taxonomic publications. **TreatmentBank (http://treatmentbank.org)** offers the ability to store and disseminate these data globally (see Figure 1). They can be chained together and configured to process automatically the input of new articles or convert entire journal runs.

Figure 1. The increased value of articles that are converted into machine-readable documents and made accessible on TreatmentBank (Plazi).



Expected advantages

1. **Access** to the **content of taxonomic publications** for immediate reuse of the data;
2. **Visualization** of content;
3. **Immediate access** to recently published data as well as legacy data;
4. **Linking of observation data** to source treatment and publication, and other external resources such as cited collections and gene sequences;
5. **Update** of **name services**;
6. **Broad distribution** of data; and
7. **Citability of sub-article elements**, where an observation record can cite the source treatment and thus the treatment can be read or mined for additional elements (e.g., traits).

PLAZI
TAKING CARE OF FREEDOM

EU BON

Applicability

GoldenGate Imagine (GGI) is an open source program that is available and maintained at http://plazi.org. It is currently a desktop application. An online version is planned to be released in the second part of 2016. GGI can be highly customized for journals to allow automatic processing with a high degree of accuracy and granularity. Currently, all the new Zootaxa articles are automatically processed and made accessible daily. This input in TreatmentBank is complemented by ingestion of semantically enhanced articles (e.g., Journals published by Pensoft). After initial training, GGI can be used and extended by users.

The following input format are possible:

- Born digital PDF (Portable Document Format) files;
- Scan-based PDF; and
- Imagine File Format (a specific format to store processed PDF files).

Potential users

Anybody with an interest in up-to-date, high-quality observation records for recently discovered, taxonomically revised or rare species, with a link to the source taxonomic treatment and publication. Furthermore, taxonomic name services such as GBIF, Catalogue of Life or the EU BON taxonomic backbone can be updated with names linked directly to the source publication.

Case study

For EU BON, Plazi is supplying the tools to data mine and extract observation records from the published literature. We have developed a system through which articles are discovered when they are published, imported, processed automatically, made accessible on TreatmentBank, and then sent to GBIF when new treatments are documented, which they will harvest and make accessible for use in EU BON. The respective treatments can be opened in GGI and further processed using a specific tool to discover and parse the observation records. They then will be resubmitted as DarwinCore Archives to GBIF or anybody who would like to receive this data.

Currently, 12,838 articles, 120,280 treatments and 38,809 observation records are available on TreatmentBank, with a daily increase of up to 100 new treatments.

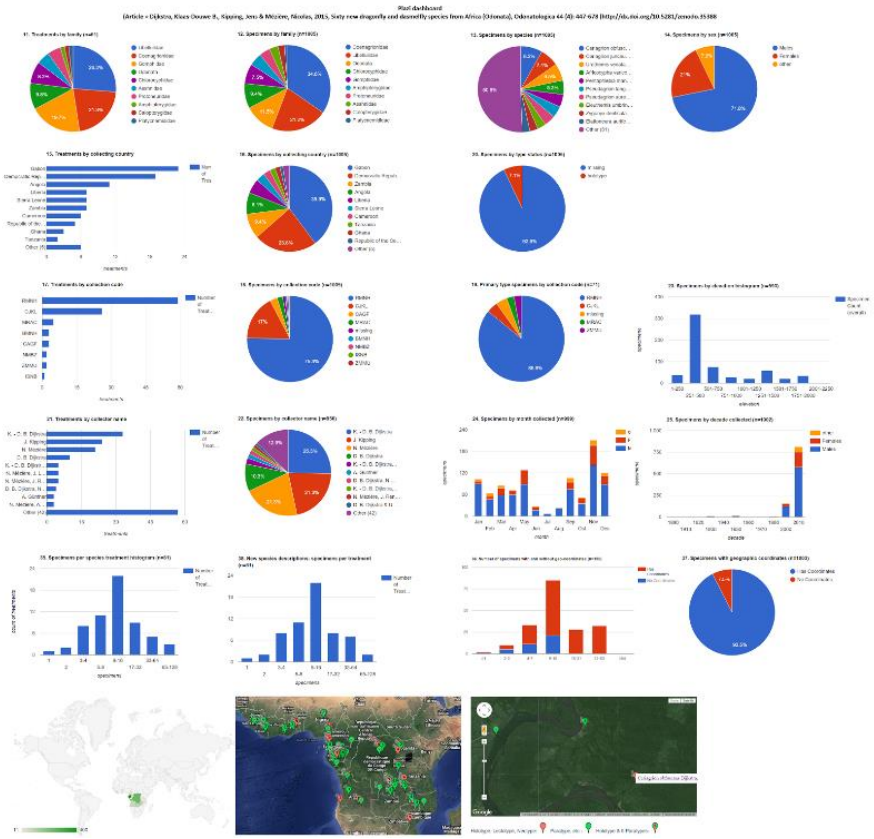


Figure 2. The visualization of the content of a taxonomic article (*Odonatologica* 44(4), 2015) illustrates the rich data contained as unstructured data within a standard scientific article.

PLAZI
TAKING CARE OF FREEDOM

EU BON