BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25748
doi: 10.3897/biss.2.25748

OPEN ACCESS

Conference Abstract

The Open Biodiversity Knowledge Management (eco-)System: Tools and Services for Extraction, Mobilization, Handling and Re-use of Data from the Published Literature

Lyubomir Penev[‡,§], Donat Agosti[|], Teodor Georgiev[‡], Viktor Senderov[‡,§], Guido Sautter[|], Terry Catapano[|,¶], Pavel Stoev[‡,#]

‡ Pensoft Publishers, Sofia, Bulgaria
§ Bulgarian Academy of Sciences, Sofia, Bulgaria
| Plazi.org, Bern, Switzerland
¶ UC Berkeley Library, Berkeley, United States of America
# National Museum of Natural History, Sofia, Bulgaria

Corresponding author: Teodor Georgiev (preprint@pensoft.net)

Received: 12 Apr 2018 | Published: 17 May 2018

Citation: Penev L, Agosti D, Georgiev T, Senderov V, Sautter G, Catapano T, Stoev P (2018) The Open Biodiversity Knowledge Management (eco-)System: Tools and Services for Extraction, Mobilization, Handling and Re-use of Data from the Published Literature. Biodiversity Information Science and Standards 2: e25748. https://doi.org/10.3897/biss.2.25748

Abstract

The Open Biodiversity Knowledge Management System (OBKMS) is an end-to-end, eXtensible Markup Language (XML)- and Linked Open Data (LOD)-based ecosystem of tools and services that encompasses the entire process of authoring, submission, review, publication, dissemination, and archiving of biodiversity literature, as well as the text mining of published biodiversity literature (Fig. 1). These capabilities lead to the creation of interoperable, computable, and reusable biodiversity data with provenance linking facts to publications.

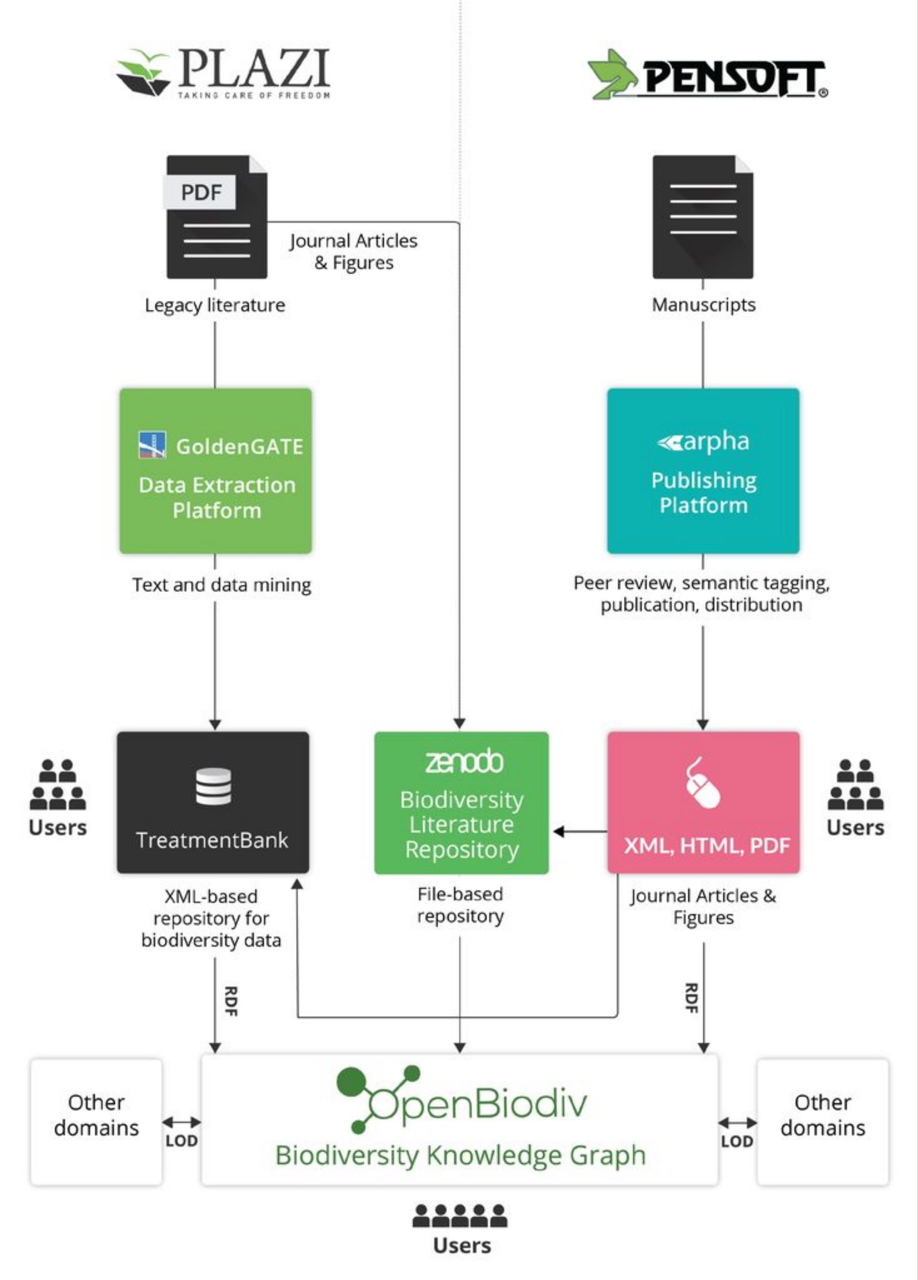

Figure 1.

The Open Biodiversity Knowledge Management (eco)-System.

OBKMS is the result of a joint endeavour by Plazi and Pensoft lasting many years. The system was developed with the support of several biodiversity informatics projects - initially

(Virtual Biodiversity Research and Access Network for Taxonomy) ViBRANT, and then followed by pro-iBiosphere, European Biodiversity Observation Network (EU BON), and Biosystematics, informatics and genomics of the big 4 insect groups (BIG4). The system includes the following key components:

1. ARPHA Journal Publishing Platform: a journal publishing platform based on the TaxPub XML extension for National Library of Medicine (NLM)'s Journal Publishing Document Type Definition (DTD) (Version 3.0). Its advanced ARPHA-BioDiv component deals with integrated biodiversity data and narrative publishing (Penev et al. 2017).
2. GoldenGATE Imagine: an environment for marking up, enhancing, and extracting text and data from PDF files, supporting the TaxonX XML schema. It has specific enhancements for articles containing descriptions of taxa ("taxonomic treatments") in the field of biological systematics, but its core features may be used for general purposes as well.
3. Biodiversity Literature repository (BLR): a public repository hosted at Zenodo (CERN) for published articles (PDF and XML) and images extracted from articles.
4. Ocellus/Zenodeo: a search interface for the images stored at BLR.
5. TreatmentBank: an XML-based repository for taxonomic treatments and data therein extracted from literature.
6. The OpenBiodiv knowledge graph: a biodiversity knowledge graph built according to the Linked Open Data (LOD) principles. Uses the RDF data model, the SPARQL Protocol and RDF Query Language (SPARQL) query language, is open to the public, and is powered by the OpenBiodiv-O ontology (Senderov et al. 2018).
7. OpenBiodiv portal:
    1. Semantic search and browser for the biodiversity knowledge graph.
    2. Multiple semantic apps packaging specific views of the biodiviersity knowledge graph.
8. Supporting tools:
    1. Pensoft Markup Tool (PMT)
    2. ARPHA Writing Tool (AWT)
    3. ReFindit
    4. R libraries for working with RDF and for converting XML to RDF (ropenbio, RDF4R).
    5. Plazi RDF converter, web services and APIs.

As part of OBKMS, Plazi and Pensoft offer the following services beyond supplying the software toolkit:

1. Digitization through imaging and text capture of paper-based or digitally born (PDF) legacy literature.
2. XML markup of both legacy and newly published literature (journals and books).
3. Data extraction and markup of taxonomic names, literature references, taxonomic treatments and organism occurrence records.
4. Export and storage of text, images, and structured data in data repositories.
5. Linking and semantic enhancement of text and data, bibliographic references, taxonomic treatments, illustrations, organism occurrences and organism traits.
6. Re-packaging of extracted information into new, user-demanded outputs via semantic apps at the OpenBiodiv portal.
7. Re-publishing of legacy literature (e.g., Flora, Fauna, and Mycota series, important biodiversity monographs, etc.).
8. Semantic open access publishing (including data publishing) of journal and books.
9. Integration of biodiversity information from legacy and newly published literature into interoperable biodiversity repositories and platforms (Global Biodiversity Information Facility (GBIF), Encyclopedia of Life (EOL), Species-ID, Plazi, Wikidata, and others).

In this presentation we make the case for why OpenBiodiv is an essential tool for advancing biodiversity science. Our argument is that through OpenBiodiv, biodiversity science makes a step towards the ideals of open science (Senderov and Penev 2016). Furthermore, by linking data from various silos, OpenBiodiv allows for the discovery of hidden facts.

A particular example of how OpenBiodiv can advance biodiversity science is demonstrated by the OpenBiodiv's solution to "taxonomic anarchy" (Garnett and Christidis 2017). "Taxonomic anarchy" is a term coined by Garnett and Christidis to denote the instability of taxonomic names as symbols for taxonomic meaning. They propose an "authoritarian" top-down approach to stablize the naming of species. OpenBiodiv, on the other hand, relies on taxonomic concepts as integrative units and therefore integration can occur through alignment of taxonomic concepts via Region Connection Calculus (RCC-5) (Franz and Peet 2009). The alignment is "democratically" created by the users of system but no consensus is forced and "anarchy" is avoided by using unambiguous taxonomic concept labels (Franz et al. 2016) in addition to Linnean names.

Presenting author

Teodor Georgiev

References

- Franz N, Chen M, Kianmajd P, Yu S, Bowers S, Weakley A, Ludäscher B (2016) Names are not good enough: Reasoning over taxonomic change in the Andropogon complex1. Semantic Web 7 (6): 645-667. https://doi.org/10.3233/sw-160220
- Franz NM, Peet RK (2009) Perspectives: Towards a language for mapping relationships among taxonomic concepts. Systematics and Biodiversity 7 (1): 5-20. https://doi.org/10.1017/s147720000800282x
- Garnett S, Christidis L (2017) Taxonomy anarchy hampers conservation. Nature 546 (7656): 25-27. https://doi.org/10.1038/546025a
- Penev L, Georgiev T, Geshev P, Demirov S, Senderov V, Kuzmova I, Kostadinova I, Peneva S, Pavel Stoev (2017) ARPHA-BioDiv: A Toolbox for Scholarly Publication and Dissemination of Biodiversity Data Based on the ARPHA Publishing Platform. Research Ideas and Outcomes 3: e13088. https://doi.org/10.3897/rio.3.e13088
- Senderov V, Penev L (2016) The Open Biodiversity Knowledge Management System in Scholarly Publishing. Research Ideas and Outcomes 2: e7757. https://doi.org/10.3897/rio.2.e7757
- Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris R, Penev L (2018) OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 9 (1): 5. https://doi.org/10.1186/s13326-017-0174-5