Progress with Repository-based Annotation Infrastructure for Biodiversity Applications

Peter Cornwell ‡

‡ Data Futures GmbH, Leipzig, Germany

Corresponding author: Peter Cornwell (peter.cornwell@data-futures.org)

Abstract

Rapid development since the 1980s of technologies for analysing texts, has led not only to widespread employment of text 'mining', but also to now-pervasive large language model artificial intelligence (AI) applications. However, building new, concise, data resources from historic, as well as contemporary scientific literature, which can be employed efficiently at scale by automation and which have long-term value for the research community, has proved more elusive.

Efforts at codifying analyses, such as the Text Encoding Initiative (TEI), date from the early 1990s and were initially driven by the social sciences and humanities (SSH) and linguistics communities, and extended with multiple XML-based tagging schemes, including in biodiversity (Miller et al. 2012). In 2010, the Bio-Ontologies Special Interest Group (of the International Society for Computational Biology) presented its Annotation Ontology (AO), incorporating JavaScript Object Notation and broadening previous XML-based approaches (Ciccarese et al. 2011). From 2011, the Open Annotation Data Model (OADM) (Sanderson et al. 2013) focused on cross-domain standards with utility for Web 3.0, leading to the W3C Web Annotation Data Model (WADM) Recommendation in February 2017*[1] and the potential for unifying the multiplicity of already-in-use tagging approaches.

This continual evolution has made the preservation of investment using annotation methods, and in particular of the connections between annotations and their context in source literature, particularly challenging. Infrastructure that entered service during the intervening years does not yet support WADM, and has only recently started to address the parallel emergence of page imagery-based standards such as the International Image Interoperability Framework (IIIF). Notably, IIIF instruments such as Mirador-2, which has been employed widely for manual creation and editing of annotations in SSH, continue to employ the now-deprecated OADM. Although multiple efforts now address combining IIIF and TEI text coordinate systems, they are currently fundamentally incompatible.

However, emerging repository technologies enable preservation of annotation investment to be accomplished comprehensively for the first time. Native IIIF support enables interactive previewing of annotations within repository graphical user interfaces and dynamic serialisation technologies provide compatibility with existing XML-based infrastructures. Repository access controls can permit experts to trace annotation sources in original texts even if the literature is not publicly accessible, e.g., due to copyright restriction. This is of paramount importance, not only because surrounding context can be crucial to qualify formal terms that have been annotated, such as collecting country. Also, contemporary automated text mining—essential for operation at the scale of known biodiversity literature—is not 100% accurate and manual checking of uncertainties is currently essential. On-going improvement of language analysis tools through AI integration offers significant future gains from reprocessing literature and updating annotation data resources. Nevertheless, without effective preservation of digitized literature, as well as annotations, this enrichment will not be possible—and today's investments in gathering together, as well as analysing scientific literature will be devalued or lost.

We report new functionality included in the InvenioRDM*[2] Free and Open Source Software (FOSS) repository software platform, which natively supports IIIF and WADM. InvenioRDM development and maintenance is funded and managed by an international consortium. From late 2023, the InvenioRDM-based ZenodoRDM update*[3] will display annotations on biodiversity literature interactively. Significantly, the Biodiversity Literature Repository (BLR) is a Zenodo Community. BLR automatically notifies the Global Biodiversity Information Facility (GBIF) of new taxonomic data and GBIF downloads and integrates this into its service.

Moreover, an annotation service based on the WADM-native Mirador-3 FOSS IIIF viewer has now been developed and will enter service with ZenodoRDM. This enables editing of biodiversity annotations from within the repository interface, as well as automated updating of taxonomic information products provided to other major infrastructures such as GBIF.

Two aspects of this ZenodoRDM annotation service are presented:

- dynamic transformation of (preservable) WADM annotations for consumption by contemporary IIIF-compliant applications such as Mirador-3, as well as for Plazi TreatmentBank/GBIF compatibility
- authentication and task organization permitting management of groups of expert contributors performing annotation enrichment tasks directly through the ZenodoRDM graphical user interface (GUI)

Workflows for editing existing biodiversity annotations, as well as origination of new annotations, need to be tailored for specific tasks—e.g., unifying geographic collecting location definitions in historic reports—via configurable dialogs for contributors and controlled vocabularies. Selectively populating workflows with annotations according to a task definition is also important to avoid cluttering the editing GUI with non-essential

information. Updated annotations are integrated into a new annotation collection upon completion of a task, before updating repository records.

Current work on annotation workflows for SSH applications is also reported. The ZenodoRDM biodiversity annotation service implements a generic repository micro-service API, and the implementation of similar services for other repository software platforms is discussed.

Keywords

biodiversity literature, IIIF, InvenioRDM, WADM, Zenodo

Presenting author

Peter Cornwell

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Ciccarese P, Ocana M, Garcia Castro LJ, et al. (2011) An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics 2 (Suppl 2). https://doi.org/10.1186/2041-1480-2-S2-S4
- Miller J, Dikow T, Agosti D, et al. (2012) From taxonomic literature to cybertaxonomic content. BMC Biol 10: 87. https://doi.org/10.1186/1741-7007-10-87
- Sanderson R, Ciccarese P, Van de Sompel H (2013) Designing the W3C open annotation data model. Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13). https://doi.org/10.1145/2464464.2464474

Endnotes

*1 https://www.w3.org/TR/annotation-model/
*2 https://inveniosoftware.org/products/rdm/
*3 https://blog.zenodo.org/2022/12/07/2022-12-07-zenodo-on-inveniordm/