Enriching Wikidata with information from OpenBiodiv about type specimens in context from different literature sources (Topic 7)

Aim/problem/goal

To develop a workflow that integrates knowledge about type materials from the OpenBiodiv knowledge graph with existing Wikidata records to enrich Wikidata with more data from biodiversity literature.

Method

The OpenBiodiv knowledge graph (Dimitrova et al. 2021) containing Linked Open Data statements extracted from literature through XML-tagging in publications was used as a data source. SPARQL queries were performed to OpenBiodiv to explore collections and institutions which have been used in the description of new taxa in the Biodiversity Data Journal (BDJ) using type specimens. Mapping of institutions and collections between OpenBiodiv, GBIF, Wikidata was done manually in some cases, when no identifier was available. Enrichment of Wikidata with information about type materials was done using OpenRefine.

Results

It was discovered that only about 2% (314 out of 14390) of all institutions and collections from the GBIF Registry of Scientific Collections (GRSciColl) are indexed in Wikidata with their GRSciColl identifier, with some being indexed without GRSciColl identifier. In addition, only 303 type specimen records were found in Wikidata. OpenBiodiv was used to discover information about type specimen locations and holding institutions/collections and map them to existing type specimen records on Wikidata, as well as create new ones, whilst keeping a reference to the original source (taxonomic article). Our contributions to Wikidata are accessible at: https://www.wikidata.org/wiki/Special:Contributions/ROMEnEwr.

Conclusion

The examination of existing records of institutions and collections in Wikidata and GRSciColl showed an ambiguous use of the term “collection”, as some institutions are regarded as collections and vice versa. A clear need to disambiguate institutions became evident, due to the multitude of identifiers used for a single institution in GRSciColl, as well as duplicate institution code records. We suggest also a revision of Wikidata properties to help capture information about institutions and type specimens in a better way.

References

- Dimitrova, Mariya, Viktor Senderov, Teodor Georgiev, Georgi Zhelezov, and Lyubomir Penev. 2021. ‘Infrastructure and Population of the OpenBiodiv Biodiversity Knowledge Graph’. *Biodiversity Data Journal* 9 (September): e67671. https://doi.org/10.3897/BDJ.9.e67671.