On a BiCIKL to Wikidata: Harmonizing the chaotic universe of natural history collectors

Mathias Dillen[‡], Andreas Plank[§], Quentin Groom[‡]

‡ Meise Botanic Garden, Meise, Belgium
§ Botanical Garden and Botanical Museum, Berlin, Germany

Corresponding author: Mathias Dillen (mathias.dillen@plantentuinmeise.be)

Abstract

People play a key role in science and have been getting increased recognition for this work, initially by the efforts of libraries to harmonize attribution of authors of creative works, such as in the Virtual International Authority File (VIAF). Now, in the realm of scientific publishing, the international ORCID (Open Researcher and Contributor ID) organization allows scientists to mint globally unique and self-maintained identifiers for themselves, which tie all of their scientific output together. This has been fairly successful for formal publications, but less so for other scientific contributions, such as samples or specimens collected from organisms to serve as vouchers or references for long-term study. Such specimens may end up in natural history collections, where their metadata and, in particular, their taxonomic identity is refined over the years by experts. As the cost of digitization has decreased and the methods become more refined, digital versions of these specimens are now being published online in vast numbers, making attribution with identifiers such as ORCID easier.

However, many of these specimens are only linked to the people who worked on them by name strings, often using a variety of syntaxes, transliterations and abbreviations. Recently, some collections have made an effort to disambiguate these names by enriching them with an explicit link to a persistent identifier, such as an ORCID (Groom et al. 2022, Little et al. 2022, von Mering et al. 2022, Meeus et al. 2023). In the European Union-funded BiCIKL (Biodiversity Community Integrated Knowledge) project, one of the goals is to develop better workflows that aid and streamline this enrichment process. This approach focuses primarily on Wikidata, as a fall-back resource for cases where ORCID identifiers are not available. Wikidata is suitable for people who are unable to register for an ORCID identifier, but also acts as a broker to harmonize between other authority sources such as the fore-mentioned VIAF.

In this presentation, we will show the results achieved so far in this BiCIKL effort. First, we will provide a landscape analysis of different person identifier enrichment efforts so far in the natural history sector. Infrastructures such as the Global Biodiversity Information Facility (GBIF), the related Bionomia and the Botany Pilot (Güntsch et al. 2021) will play a

key role. Building on those results, we will identify improved methods to link name strings to unique biographical records in resources such as Wikidata and ORCID. The focus will primarily be on properties of these records that make them more identifiable as collectors of specimens, and that can systematically be harnessed to rank multiple possible candidates or reject false positives. Finally, these findings will be incorporated into workflows that streamline the disambiguation and enrichment process. These workflows will build on existing pipelines and platforms, such as Bionomia and the Botany Pilot, focusing in particular on data roundtripping from these resources into other infrastructures, such as local collection management systems, but also specimen data aggregators such as GBIF and the Distributed System of Scientific Collections (DiSSCo).

Keywords

semantic enrichment, PIDs, matching, roundtripping, Bionomia

Presenting author

Mathias Dillen

Presented at

TDWG 2023

Acknowledgements

This work was done under the BiCIKL project, using funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Groom Q, Bräuchler C, Cubey R, Dillen M, Huybrechts P, Kearney N, Klazenga N, Leachman S, Paul DL, Rogers H, Santos J, Shorthouse D, Vaughan A, von Mering S, Haston E (2022) The disambiguation of people names in biological collections. Biodiversity Data Journal 10 https://doi.org/10.3897/bdj.10.e86089
- Güntsch A, Groom Q, Ernst M, Holetschek J, Plank A, Röpert D, Fichtmüller D, Shorthouse DP, Hyam R, Dillen M, Trekels M, Haston E, Rainer H (2021) A botanical demonstration of the potential of linking data using unique identifiers for people. PLOS ONE 16 (12). https://doi.org/10.1371/journal.pone.0261130

- Little H, Karim T, Krimmel E, Norton B, Utrup J, Walker L, Van Veldhuizen J (2022) Community Data Mobilization in Wikidata: A paleontology perspective. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.94416
- Meeus S, Ariño A, Bakken T, Braun P, Dillen M, Endresen D, Haston EM, Lowe M, Meyke E, Santos J, von Mering S, Groom Q (2023) Who is Who in Natural History Collections? Zenodo https://doi.org/10.5281/zenodo.7781754
- von Mering S, Kaiser K, Petersen M (2022) Transforming Closed Silos into Shared Resources: Opening up data on historical collection agents affiliated with the Museum für Naturkunde Berlin. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.93787