Is Your Collection Ambiguous?

Mathias Dillen[‡], Elspeth M Haston[§], Nicole Kearney[|,¶,#], Deborah L Paul[¤,«], Joaquim Santos[»], David Peter Shorthouse[^], Alison Vaughan[˅], Sabine von Mering[¦], Quentin Groom[‡]

‡ Meise Botanic Garden, Meise, Belgium
§ Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom
| Biodiversity Heritage Library (BHL), Melbourne, Australia
¶ Atlas of Living Australia (ALA), Melbourne, Australia
# Museums Victoria (MV), Melbourne, Australia
¤ University of Illinois Urbana-Champaign, Urbana, United States of America
« Florida State University, Tallahassee, United States of America
» Herbarium of the University of Coimbra, Coimbra, Portugal
^ Agriculture & Agri-Food Canada, Ottawa, Canada
˅ Royal Botanic Gardens Victoria, Melbourne, Australia
¦ Museum für Naturkunde, Berlin, Germany

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

Abstract

The natural history specimens of the world have been documented on paper labels, often physically attached to the specimen itself. As we transcribe these data to make them digital and more useful for analysis, we make interpretations. Sometimes these interpretations are trivial, because the label is unambiguous, but often the meaning is not so clear, even if it is easily read. One key element that suffers from considerable ambiguity is people's names. Though a person is indivisible, their name can change, is rarely unique and can be written in many ways. Yet knowing the people associated with data is incredibly useful. Data on people can be used to validate other data, simplify data capture, link together data across domains, reduce duplication-of-effort and facilitate data-gap-analysis. In addition, people data enable the discovery of individuals unique to our collections, the collective charting of the history of scientific researchers and the provision of credit to the people who deserve it (Groom et al. 2020).

We foresee a future where the people associated with collections are not ambiguous, are shared globally, and data of all kinds are linked through the people who generate them. The TDWG People in Biodiversity Data Task Group is therefore working on a guide to the disambiguation of people in natural history collections. The ultimate goal is to connect the various strings of characters on specimen labels and other documentation to persistent identifiers (PIDs) that unambiguously link a name "string" to the identity of a person. In working towards this goal, 150 volunteers in the Bionomia project have linked 21 million specimens to persistent identifiers for their collectors and determiners. An additional 2 million specimens with links to identifiers for people have already emerged directly from collections that make use of the recently ratified Darwin Core terms recordedByID and identifiedByID. Furthermore, the CETAF Botany Pilot conducted among a group of European

herbaria and museums has connected over 1.4 million specimens to disambiguated collectors (Güntsch et al. 2021). Still, given the estimated 2 billion (Ariño 2010) natural history specimens globally, there is much more disambiguation to be done.

The process of disambiguation starts with a trigger, which is often the transcription of a specimen's label data. Unambiguous identification of the collector may facilitate this transcription, as it offers knowledge of their biographical details and collecting habits, allowing us to infer missing information such as collecting date or locality. Another trigger might be the flagging of inconsistent data during data entry or resulting from data quality processes, revealing for instance that multiple collectors have been conflated. A disambiguation trigger is followed by the gathering of data, then the evaluation of the results and finally by the documentation of the new information.

Disambiguation is not always straightforward and there are many pitfalls. It requires access to biographical data, and identifiers to be minted. In the case of living people, they have to cooperate with being disambiguated and we have to follow legal and ethical guidelines. In the case of dead people, particularly those long dead, disambiguation may require considerable research.

We will present the progress made by the People in Biodiversity Data Task Group and their recommendations for disambiguation in collections. We want to encourage other institutions to engage with a global effort of linking people to persistent identifiers to collaboratively improve all collection data.

Keywords

disambiguation, Wikidata, PID, agents, interoperability

Presenting author

Mathias Dillen

Presented at

TDWG 2021

Conflicts of interest

References

- Ariño A (2010) Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7 (2). https://doi.org/10.17161/bi.v7i2.3991

- Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen AE, Haston E (2020) People are essential to linking biodiversity data. Database 2020 https://doi.org/10.1093/database/baaa072
- 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. Manuscript submitted for publication.