Progress in Authority Management of People Names for Collections

Quentin J. Groom[‡], Chloé Besombes[§], Josh Brown[|], Simon Chagnoux[§], Teodor Georgiev[¶], Nicole Kearney[#], Arnald Marcer[¤], Nicky Nicolson[«], Roderic Page[»], Sarah Phillips[^], Heimo Rainer[ˇ], Greg Riccardi[¦], Dominik Röpert[ʔ], David Peter Shorthouse[ˤ], Pavel Stoev[¢], Elspeth Margaret Haston[ℓ]

‡ Meise Botanic Garden, Meise, Belgium
§ Muséum National d'Histoire Naturelle, Paris, France
| ORCID, Bethesda, United States of America
¶ Pensoft Publishers, Sofia, Bulgaria
# Biodiversity Heritage Library Australia, Melbourne, Australia
¤ CREAF, E08193 Bellaterra (Cerdanyola del Vallès), Catalonia, Spain
« Royal Botanic Gardens, Kew, Kew, United Kingdom
» University of Glasgow, Glasgow, United Kingdom
^ Royal Botanic Gardens Kew, Richmond, United Kingdom
ˇ Naturhistorisches Museum Wien, Wien, Austria
¦ Florida State University, Tallahassee, United States of America
ʔ Botanic Garden and Botanical Museum Berlin, Berlin, Germany
ˤ Agriculture and Agri-Food Canada, Ottawa, Canada
¢ National Museum of Natural History and Pensoft Publishers, Sofia, Bulgaria
ℓ Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom

Corresponding author: Quentin J. Groom (quentin.groom@plantentuinmeise.be)

Abstract

The concept of building a network of relationships between entities, a knowledge graph, is one of the most effective methods to understand the relations between data. By organizing data, we facilitate the discovery of complex patterns not otherwise evident in the raw data.

Each datum at the nodes of a knowledge graph needs a persistent identifier (PID) to reference it unambiguously. In the biodiversity knowledge graph, people are key elements (Page 2016). They collect and identify specimens, they publish, observe, work with each

other and they name organisms. Yet biodiversity informatics has been slow to adopt PIDs for people and people are currently represented in collection management systems as text strings in various formats. These text strings often do not separate individuals within a collecting team and little biographical information is collected to disambiguate collectors.

In March 2019 we organised an international workshop to find solutions to the problem of PIDs for people in collections with the aim of identifying people unambiguously across the world's natural history collections in all of their various roles. Stakeholders were represented from 11 countries, representing libraries, collections, publishers, developers and name registers.

We want to identify people for many reasons. Cross-validation of information about a specimen with biographical information on the specimen can be used to clean data. Mapping specimens from individual collectors across multiple herbaria can geolocate specimens accurately. By linking literature to specimens through their authors and collectors we can create collaboration networks leading to a much better understanding of the scientific contribution of collectors and their institutions. For taxonomists, it will be easier to identify nomenclatural type and syntype material, essential for reliable typification. Overall, it will mean that geographically dispersed specimens can be treated much more like a single distributed infrastructure of specimens as is envisaged in the European Distributed Systems of Scientific Collections Infrastructure (DiSSCo).

There are several person identifier systems in use. For example, the Virtual International Authority File (VIAF) is a widely used system for published authors. The International Standard Name Identifier (ISNI), has broader scope and incorporates VIAF. The ORCID identifier system provides self-registration of living researchers. Also, Wikidata has identifiers of people, which have the advantage of being easy to add to and correct. There are also national systems, such as the French and German authority files, and considerable sharing of identifiers, particularly on Wikidata. This creates an integrated network of identifiers that could act as a brokerage system. Attendees agreed that no one identifier system should be recommended, however, some are more appropriate for particular circumstances.

Some difficulties have still to be resolved to use those identifier schemes for biodiversity : 1) duplicate entries in the same identifier system; 2) handling collector teams and preserving the order of collectors; 3) how we integrate identifiers with standards such as Darwin Core, ABCD and in the Global Biodiversity Information Facility; and 4) many living and dead collectors are only known from their specimens and so they may not pass notability standards required by many authority systems.

The participants of the workshop are now working on a number of fronts to make progress on the adoption of PIDs for people in collections. This includes extending pilots that have already been trialed, working with identifier systems to make them more suitable for specimen collectors and talking to service providers to encourage them to use ORCID iDs to identify their users. It was concluded that resolving the problem of person

identifiers for collections is largely not a lack of a solution, but a need to implement solutions that already exist.

Keywords

authority file, identifer, VIAF, Wikidata, PID, ORCID

Presenting author

Quentin Groom

Grant title

References

- Page R (2016) Towards a biodiversity knowledge graph. Research Ideas and Outcomes https://doi.org/10.3897/rio.2.e8767