Harmonizing person names in Biodiversity Informatics using Wikidata

Mathias Dillen[1], Quentin Groom[1], Andreas Plank[2]

1 : Meise Botanic Garden, BE
2 : Botanic Garden Berlin, DE

People in biodiversity?

1. People do all the work!
2. Hence they tie together data from different domains and disciplines
3. By linking up data, we can do more with it, fill in gaps, validate potential errors
4. Person names are hard to parse: # syntaxes, transliterations, abbreviations
5. Use Persistent Identifiers instead: ORCID, VIAF, Wikidata

Landscape of attributions

Specimens:

GBIF
Bionomia
Botany Pilot

Table 2: *Non-observation datasets with most attributions.*

Dataset Key	# Records	%	% cumulativ	Dataset Name
b740eaa0-0679-41dc-acb7-990d562dfa37	1,602,826	28.5%	28.5%	Meise Botanic Garden Herbarium (BR)
e45c7d91-81c6-4455-86e3-2965a5739b1f	795,815	14.2%	42.7%	Vascular Plant Herbarium, Oslo (O) UiO
4ce8e3f9-2546-4af1-b28d-e2eadf05dfd4	572,097	10.1%	52.8%	National Herbarium of Victoria (MEL) AVH data
4bfac3ea-8763-4f4b-a71a-76a6f5f243d3	508,590	9.1%	61.9%	Museum of Comparative Zoology, Harvard University
26f5b360-8770-4d54-9c2d-397798a5e513	386,778	6.9%	68.8%	Entomology, Oslo (O) UiO
ffae417e-b2d8-476c-afe4-8c1093b67071	375,496	6.6%	75.4%	Bacterial members of the Pinus pinaster rhizosphere microbiota in a forest subjected to drought conditions
7948250c-6958-4a29-a670-ed1015b26252	243,946	4.4%	79.8%	Lichen herbarium, Oslo (O) UiO
e4deab67-0998-4140-b573-0ba1f624eb3e	202,874	3.6%	83.4%	Fungarium, Oslo (O) UiO
68a0650f-96ae-499c-8b2a-a4f92c01e4b3	174,819	3.1%	86.5%	Bryophyte Herbarium, Oslo (O) UiO
2b044aa9-1a9a-413e-8b18-ed09da575d3f	95,190	1.7%	88.2%	University of Tartu Natural History Museum and Botanical Garden Zoological Collections
a559e942-0fbe-4f09-93ca-28cf244ce2a0	79,674	1.4%	89.6%	Herbarium of University of Coimbra (COI)

Improved methods of finding people in WIKIDATA

Optimize SPARQL queries

Enrich Wikidata Content

SPARQL filters used to find likely collectors/determiners in Wikidata.

Property	property id	item id
occupation: botanist	P106/P279*	Q2374149
occupation: zoologist	P106/P279*	Q350979
Bionomia ID	P6944	
Harvard Index of Botanists ID	P6264	
IPNI author ID	P586	
BHL creator ID	P4081	
Entomologists of the World ID	P5370	
Zoobank author ID	P2006	
collection items at Wikispecies	P11146	

Automated matching of people names to Wikidata

R and Python scripts to automatically match names to Wikidata records

- Rule based
- Clustering based

Pilot results

Variable, up to 50% matchable

1. Missing in Wikidata
2. Not findable in Wikidata
3. Not just names
4. Not specific enough

	Naturalis Botany	Virtual Herbarium Germany	Meise Botanic Garden Herbarium
parsed names	58.815	27.613	67.696
specimens	4.991.356	1.098.627	2.779.376
unique strings	100.857	64.440	85.774
single match	14.286	9.943	13.462
multi match	7.286	3.929	6.350
no match	37.243	12.041	43.855
quick statements	12.479	11.829	9.865
new claims	10.444	10.962	7.947

Roundtripping the attributions?

- DiSSCo annotations
- Darwin Core attributions (deprecated)
- Wikidata (but not at the specimen level)
- Nanopublications?
- CETAF Identifier RDF metadata

Read more

- https://github.com/AgentschapPlantentuinMeise/collector_matching
- https://github.com/infinite-dao/collector-matching/
- https://doi.org/10.3897/arphapreprints.e114920
- https://doi.org/10.3897/BDJ.10.e86089

This poster:

This project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101007492