

Conference Abstract

Linking Data and Descriptions on Moths Using the Wikimedia Ecosystem

Andra Waagmeester[‡], Paul Braun[§], Manoj Karingamadathil^{||}, Jose Emilio Labra Gayo[¶], Siobhan Leachman[#], Katherine Thornton[□]

[‡] Micelio, Ekeren (Antwerp), Belgium

[§] Musée national d'histoire naturelle Luxembourg, Luxembourg, Luxembourg

| Kerala Biodiversity Monitoring Network, Thrissur, India

[¶] WESO - University of Oviedo, Oviedo, Spain

[#] Citizen Scientist, Wellington & Wairarapa, New Zealand

[□] Citizen Scientist, Olympia, United States of America

Corresponding author: Andra Waagmeester (andra@micelio.be)

Received: 01 Sep 2021 | Published: 03 Sep 2021

Citation: Waagmeester A, Braun P, Karingamadathil M, Labra Gayo JE, Leachman S, Thornton K (2021) Linking Data and Descriptions on Moths Using the Wikimedia Ecosystem. Biodiversity Information Science and Standards 5: e73806. <https://doi.org/10.3897/biss.5.73806>

Abstract

Moths form a diverse group of species that are predominantly active at night. They are colourful, have an ecological role, but are less well described compared to their closest relatives, the butterflies. Much remains to be understood about moths, which is shown by the many issues within their taxonomy, including being a paraphyletic group and the inability to clearly distinguish them from butterflies (Fig. 1).

We present the Wikimedia architecture as a hub of knowledge on moths. This ecosystem consists of [312](#) language editions of [Wikipedia](#) and sister projects such as Wikimedia [com mons](#) (a multimedia repository), and [Wikidata](#) (a public knowledge graph).

Through Wikidata, external data repositories can be integrated into this knowledge landscape on moths. Wikidata contains links to (open) data repositories on biodiversity like [iNaturalist](#), Global Biodiversity Information Facility ([GBIF](#)) and the Biodiversity Heritage Library ([BHL](#)) which in return contain detailed content like species occurrence data, images or publications on moths.

al. 2021). These schemas can then be used to verify if a subset of Wikidata conforms to an expected or described data model.

Starting from a document that describes an expected schema on moths, we have developed an EntitySchema ([E321](#)) for moths in Wikidata. This schema provides unambiguous guidance for contributors who have data they are not sure how to model. For example, a user with data about a particular species of moth may be working from a scientific article that states that the species is only found in New Zealand, and may be unsure of how to model that fact as a statement in Wikidata. After consulting Schema E321, the user will find out about Property P183 “endemic_to” and then use that property to state that the species is endemic to New Zealand. As more contributors follow the data model expressed in schema E321, there will be structural consistency across items for moths in Wikidata. This reduces the risk of contributors using different combinations of properties and qualifiers to express the same meaning. If a contributor needs to express something that is not yet represented in Schema E321 they can extend the schema itself, as each schema can be edited. The multilingual affordances of the Wikidata platform allow users to edit in over 300 languages. In this way, contributors edit in their preferred language and see the structure of the data as well as the schemas in their language of choice. This broadens the range of people who can contribute to these data models and reduces the dominance of English.

There are approximately [160K+](#) estimated moth species. This number is equal to the number of moths described in iNaturalist, while Wikidata contains 220K items on moths. As the biggest language edition, the English Wikipedia contains 65K moth articles; other language editions contain far fewer Wikipedia articles. The higher number of items on moths in Wikidata can be partly explained by Wikidata taxon synonyms being treated as distinct taxa.

Wikidata, as a proxy of knowledge on moths, is instrumental in getting them better described in Wikipedia and other (FAIR) sources. While in return, curation in Wikidata happens by a large community. This approach to data modelling has the advantage of allowing multilingual collaboration and iterative extension and improvement over time.

Keywords

Wikidata, data schemas

Presenting author

Andra Waagmeester

Presented at

TDWG 2021

Acknowledgements

This work builds on the passionate efforts in the biodiversity community, specifically in iNaturalist and all the connected communities through GBIF. Also a big shoutout to the connected communities active in the different Wikimedia sister projects, most notably Wikidata.

References

- Thornton K, Solbrig H, Stupp G, Labra Gayo JE, Mietchen D, Prud'hommeaux E, Waagmeester A (2019) Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation. The Semantic Web606-620. https://doi.org/10.1007/978-3-030-21348-0_39
- Waagmeester A, Willighagen EL, Su AI, Kutmon M, Gayo JEL, Fernández-Álvarez D, Groom Q, Schaap PJ, Verhagen LM, Koehorst JJ (2021) A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses. BMC biology 19 (1): 12. <https://doi.org/10.1186/s12915-020-00940-y>
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). <https://doi.org/10.1038/sdata.2016.18>