BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59046
doi: 10.3897/biss.4.59046

OPEN ACCESS

Conference Abstract

# Unleash the Potential of your Website! 180,000 webpages from the French Natural History Museum marked up with Bioschemas/Schema.org biodiversity types

Franck Michel[‡], Gargominy Olivier[§], Benjamin Ledentec[§], The Bioschemas Community[|]

‡ Université Côte d'Azur, CNRS, Inria, I3S, Sophia-Antipolis, France
§ Muséum national d'Histoire naturelle, Paris, France
| Multiple affiliations, n/a, United Kingdom

Corresponding author: Franck Michel (franck.michel@cnrs.fr)

Received: 28 Sep 2020 | Published: 29 Sep 2020

Citation: Michel F, Olivier G, Ledentec B, Community TB (2020) Unleash the Potential of your Website! 180,000 webpages from the French Natural History Museum marked up with Bioschemas/Schema.org biodiversity types. Biodiversity Information Science and Standards 4: e59046. https://doi.org/10.3897/biss.4.59046

## Abstract

The challenge of finding, retrieving and making sense of biodiversity data is being tackled by many different approaches. Projects like the Global Biodiversity Information Facility (GBIF) or Encyclopedia of Life (EoL) adopt an integrative approach where they republish, in a uniform manner, records aggregated from multiple data sources. With this centralized, siloed approach, such projects stand as powerful one-stop shops, but tend to reduce the visibility of other data sources that are not (yet) aggregated. At the other end of the spectrum, the Web of Data promotes the building of a global, distributed knowledge graph consisting of datasets published by independent institutions according to the Linked Open Data principles (Heath and Bizer 2011), such as Wikidata or DBpedia. Beyond these "sophisticated" infrastructures, websites remain the most common way of publishing and sharing scientific data at low cost. Thanks to web search engines, everyone can discover webpages. Yet, the summaries provided in results lists are often insufficiently informative to decide whether a web page is relevant with respect to some research interests, such that integrating data published by a wealth of websites is hardly possible. A strategy around this issue lies in annotating websites with structured, semantic metadata such as the

Schema.org vocabulary (Guha et al. 2015). Webpages typically embed Schema.org annotations in the form of markup data (written in the RDFa or JSON-LD formats), which search engines harvest and exploit to improve ranking and provide more informative summarization.

Bioschemas is a community effort working to extend Schema.org to support markup for Life Sciences websites (Michel and The Bioschemas Community 2018, Garcia et al. 2017). Bioschemas primarily re-uses existing terms from Schema.org, occasionally re-uses terms from third-party vocabularies, and when necessary proposes new terms to be endorsed by Schema.org. As of today, Bioschemas's biodiversity group has proposed the *Taxon* type*1 to support the annotation of any webpage denoting taxa, *TaxonName* to support more specifically the annotation of taxonomic names registries, and guidelines describing how to leverage existing vocabularies such as Darwin Core terms.

To proceed further, the biodiversity community must now demonstrate its interest in having these terms endorsed by Schema.org: (1) through a critical mass of live markup deployments, and (2) by the development of applications capable of exploiting this markup data.

Therefore, as a first step, the French National Museum of Natural History has marked up its natural heritage inventory website: over 180,000 webpages describing the species inventoried in French territories have been annotated with the *Taxon* and *TaxonName* types in the form of JSON-LD scripts (see example scripts). As an example, one can check the source of the Delphinus delphis page.

In this presentation, by demonstrating that marking up existing webpages can be very inexpensive, we wish to encourage the biodiversity community to adopt this practice, engage in the discussion about biodiversity-related markup, and possibly propose new terms related e.g. to traits or collections. We believe that generalizing the use of such markup by the many websites reporting checklists, museum collections, occurrences, life traits etc. shall be a major step towards the generalized adoption of FAIR*2 principles (Wilkinson 2016), shall dramatically improve information discovery using search engines, and shall be a key accelerator for the development of novel, web-scale, biodiversity data integration scenarios.

## Presenting author

Franck Michel

## Presented at

TDWG 2020

## References

- Garcia L, Giraldo O, Garcia A, Dumontier M, Bioschemas Community (2017) Bioschemas: schema. org for the Life Sciences. 2042. Proceedings of the Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4LS).
- Guha RV, Brickley D, MacBeth S (2015) Schema.org: Evolution of Structured Data on the Web. ACM Queue - Structured Data 13 (9). https://doi.org/10.1145/2857274.2857276
- Heath T, Bizer C (2011) Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology 1 (1). https://doi.org/10.2200/S00334ED1V01Y201102WBE001
- Michel F, The Bioschemas Community, et al. (2018) Bioschemas & Schema.org: a Lightweight Semantic Layer for Life Sciences Websites. Biodiversity Information Science and Standards, TDWG 2018 Proceedings. https://doi.org/10.3897/biss.2.25836
- Wilkinson MD, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3.

## Endnotes

[*1] All Bioschemas types are available at https://bioschemas.org/types/

[*2] Findable, Accessible, Interoperable, Reusable