Bioschemas & Schema.org: a Lightweight Semantic Layer for Life Sciences Websites

Franck Michel

BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25836
doi: 10.3897/biss.2.25836

OPEN ACCESS

Conference Abstract

Bioschemas & Schema.org: a Lightweight Semantic Layer for Life Sciences Websites

Franck Michel‡, The Bioschemas Community§

‡ Université Côte d'Azur, CNRS, Inria, I3S, Sophia-Antipolis, France
§ Multiple affiliations, , United Kingdom

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

Received: 14 Apr 2018 | Published: 22 May 2018

Citation: Michel F, The Bioschemas Community (2018) Bioschemas & Schema.org: a Lightweight Semantic Layer for Life Sciences Websites. Biodiversity Information Science and Standards 2: e25836.
https://doi.org/10.3897/biss.2.25836

Abstract

Web portals are commonly used to expose and share scientific data. They enable end users to find, organize and obtain data relevant to their interests. With the continuous growth of data across all science domains, researchers commonly find themselves overwhelmed as finding, retrieving and making sense of data becomes increasingly difficult. Search engines can help find relevant websites, but the short summarizations they provide in results lists are often little informative on how relevant a website is with respect to research interests.

To yield better results, a strategy adopted by Google, Yahoo, Yandex and Bing involves consuming structured content that they extract from websites. Towards this end, the schem a.org collaborative community defines vocabularies covering common entities and relationships (e.g., events, organizations, creative works) (Guha et al. 2016). Websites can leverage these vocabularies to embed semantic annotations within web pages, in the form of markup using standard formats. Search engines, in turn, exploit semantic markup to enhance the ranking of most relevant resources while providing more informative and accurate summarization. Additionally, adding such rich metadata is a step forward to make data FAIR, i.e. Findable, Accessible, Interoperable and Reusable.

Although schema.org encompasses terms related to data repositories, datasets, citations, events, etc., it lacks specialized terms for modeling research entities. The Bioschemas community (Garcia et al. 2017) aims to extend schema.org to support markup for Life Sciences websites. A major pillar lies in reusing types from schema.org as well as well-adopted domain ontologies, while only proposing a limited set of new types. The goal is to enable semantic cross-linking between knowledge graphs extracted from marked-up websites. An overview of the main types is presented in Fig. 1. Bioschemas also provides profiles that specify how to describe an entity of some type. For instance, the protein profile requires a unique identifier, recommends to list transcribed genes and associated diseases, and points to recommended terms from the Protein Ontology and Semantic Science Integrated Ontology.

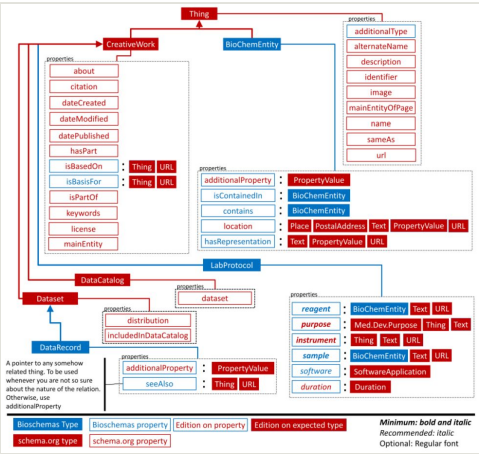

Figure 1.

Bioschemas types and properties at a glance.

The success of schema.org lies in its simplicity and the support by major search engines. By extending schema.org, Bioschemas enables life sciences research communities to benefit from a lightweight semantic layer on websites and thus facilitates discoverability and interoperability across them. From an initial pilot including just a few bio-types such as proteins and samples, the Bioschemas community has grown and is now opening up towards other disciplines. The biodiversity domain is a promising candidate for such further extensions. We can think of additional profiles to account for biodiversity-related information. For instance, since taxonomic registers are the backbone of many web portals and databases, new profiles could describe taxa and scientific names while reusing well-adopted vocabularies such as Darwin Core terms (Baskauf et al. 2016) or TDWG ontologies (TDWG Vocabulary Management Task Group 2013). Fostering the use of such markup by web portals reporting traits, observations or museum collections could not only improve information discovery using search engines, but could also be a key to spur large-scale biodiversity data integration scenarios.

Presenting author

Franck Michel

References

- Baskauf S, Wieczorek J, Deck J, Webb C (2016) Lessons Learned from Adapting the Darwin Core Vocabulary Standard for Use in RDF. Semantic Web 7 (6): 617-627.
- Garcia L, Giraldo O, Garcia A, Dumontier M, Bioschemas Community (2017) Bioschemas: schema.org for the Life Sciences. Proceedings of SWAT4LS. CEUR, 2042
- Guha R, Brickley D, MacBeth S (2016) Schema.org: Evolution of Structured Data on the Web. Communications of the ACM 59 (2): 44-51.
- TDWG Vocabulary Management Task Group (2013) Report of the TDWG Vocabulary Management Task Group (VoMaG). http://www.gbif.org/document/80862