Bioschemas: schema.org for the Life Sciences

Leyla Garcia[1[0000-0003-3986-0510]], Olga Giraldo[2[0000-0003-2978-8922]] , Alexander Garcia[2[0000-0003-1238-2539]] , Michel Dumontier[3[0000-0003-4727-9435]] and Bioschemas Community

[1] European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD, UK.
ljgarcia@ebi.ac.uk

[2] Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Boadilla del Monte, Madrid, Spain
ogiraldo@fi.upm.es, agarcia@gmail.com

[3] University of Maastricht , Minderbroedersberg 4-6, 6211 LK Maastricht, The Netherlands
michel.dumontier@maastrichtuniversity.nl

Abstract. Websites are commonly used to expose data to end users, enabling search, filter, and download capabilities making it easier for users to find, organize and obtain data relevant to their own interests. With the continuous growth of data in the Life Sciences domain, it becomes difficult for users to easily find information required for their research on one single website. Search engines should make it easier for researchers to search and retrieve collated information from multiple sites so they can better decide where to go next. Schema.org is a collaborative project providing schemas for semantically structuring data in web pages. By adding semantic mark-up it becomes easier to determine whether a web page refers to a book or a movie. It also facilitates summarizing information in a fashion similar to infoboxes used in Wikipedia. Bioschemas is a community effort aiming to extend schema.org to support mark-up for Life Sciences websites. Here we present an overview of the main types used and proposed by Bioschemas in order to support such mark up. Availability: http://bioschemas.org/

Keywords: Semantic mark-up, structured data, data discoverability.

1 Bioschemas

Bioschemas is a community initiative aiming to extend schema.org in order to improve data discoverability and interoperability in Life Sciences. Bioschemas reuses some existing types such as *DataCatalog* and *Dataset*, adds new properties to others such as *CreativeWork*, and proposes new types such as *BioChemEntity*, *DataRecord* and *LabProtocol*. Editions and additions are expected to be included in schema.org during 2018. In addition to types and properties, Bioschemas also provides guidelines regarding cardinality –one or many, marginality –minimum, recommended or optional, and usage of controlled vocabularies for those properties considered more relevant for Life

Sciences data. Specifications and guidelines are available at http://bioschemas.org/specifications. An overview of the main types involved in Bioschemas is presented in Fig. 1.

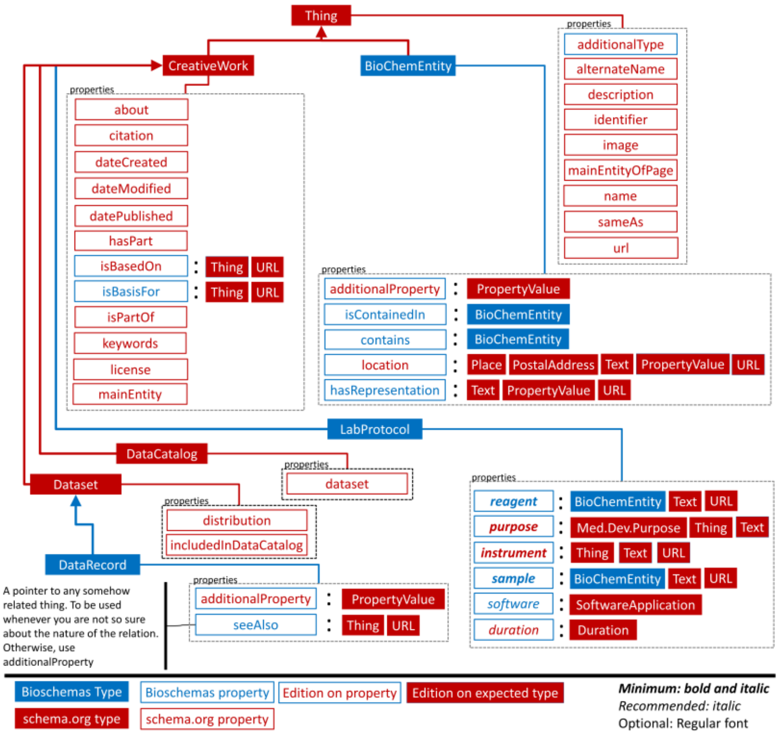

Fig. 1. Bioschemas types and properties at a glance.

BioChemEntity acts as a flexible and extensible wrapper, easy to customize. For such customizations, a.k.a. profiles, Bioschemas provides guidelines on the (i) minimum and recommended data to be delivered, (ii) expected cardinality, and (iii) third-party ontology terms useful to model the data. For instance, a protein profile advises as minimum one unique identifier while as recommended transcribed genes, organisms and associated diseases. On top of it, the protein profile recommends a well-known ontology class or controlled vocabulary type such as *http://purl.obolibrary.org/obo/PR_000000001* to represent the protein type, as well as object properties or predicates such as *http://semanticscience.org/resource/SIO_010081* to link to transcribed genes, *schema:isContainedIn* to link to organisms, and *http://semanticscience.org/resource/SIO_000001* to link to associated diseases. The property *schema:mainEntityOfPage* is used to link the entity to its corresponding *schema:DataRecord* in a *schema:Dataset*, while *schema:sameAs* is used to link to other pages describing this entity, and *schema:url* is used to link to its official webpage. Following the specifications, Bioschemas will continue with adoption by some key resources in Life Sciences and development of tools for validation and data extraction.