Bioschemas, a community initiative to extend schemas.org to the Life Sciences domain

Leyla Garcia[1], Niall Beard[2], Tony Burdett[1], Carole Goble[2], Simon Jupp[1], Nick Jutty[1],
Helen Parkinson[1], Maria Martin[1], Rafael Jiménez[3]

[1] EMBL-EBI, Wellcome Genome Campus, CB10 1 SD, Hinxton, UK
`{ljgarcia,tburdett,jupp,juty,parkinson,tmartin}ebi.ac.uk`
[2] University of Manchester, Oxford Rd, Mancheser, UK
`{niall.beard,carole.goble}`@manchester.ac.uk
3 ELIXIR, Wellcome Genome Campus, CB10 1 SD, Hinxton, UK
`rafael.jimenez@elixir-europe.org`

Abstract. Websites are commonly used to expose data to end-users, enabling search, filter, and download capabilities so users can easily find, organize and obtain data relevant to their own interests. With the continuous and distributed growth of data in the Life Sciences domain, it has become more and more difficult for users to find all the information required for their research on one single website. Effective search engines are therefore a key resource for researchers. Schema.org is a collaborative and community effort to provide schemas for semantically structuring data contained on web pages. By adding semantic mark-up it becomes easier to determine whether a web page refers to a book or a movie. Schema.org semantic mark-up is implemented through an HTML-like tag declaring, for instance, the intended type of a section on a web page. Here we present Bioschemas, a community initiative aiming to extend Schema.org to encompass support for specialized Life Sciences terms. Bioschemas aim not only to define and promote the use of such schemas but also to provide a discovery platform for users to easily access marked-up data from those websites using the proposed schemas.

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

1 Introduction

Websites are commonly used to expose data to end-users. They enable search, filter, and download capabilities so users can easily find, organize and obtain data relevant to their own interests. With the continuous growth of data in the Life Sciences (LS) domain, it becomes more and more difficult for users to find all the information required by their research on one single website. In order to facilitate users to navigate across different but related types of data, LS websites commonly link to each other. For instance, while reading about a human protein at the Universal Protein Resource (UniProt) [1] website, it is possible to navigate to ChEMBL [2] in order to find out more information about chemical aspects related to that protein.

Although linking to each other improves cross-navigation, websites still need to be findable by themselves. The higher they come on a search engine results list the better, as they will gain visibility and will thus reach more users. However, coming on top of a result list becomes useful only when the search terms are indeed related to a website. Search engines need to be clever when retrieving search results, and, in order to do so, search engines such as Google, Yahoo, Yandex and Bing use different strategies, one of them being the incorporation of structured content extracted from websites.

Schema.org is a collaborative community effort to provide schemas for structuring data on Internet [3]. Such schemas are put together in the form of a vocabulary covering common entities and relationships such as event, organization, business, and so on. The easiest way to introduce these schemas on a website is using markup. A markup looks like a Hypertext Text Markup Language (HTML) tag but rather than being used to define how the content should be displayed on a web browser, it provides information about the content type. For instance, while *<h1>100 years of solitude</h1>* will apply a heading style to the text, *<h1 itemtype="http://schema.org/Book">100 years of solitude</h1>* will also semantically state that this item refers to a book.

Here we introduce Bioschemas, a community initiative aiming to extend Schema.org to the LS domain. Bioschemas intends not only to define and promote the use of such schemas but also to provide markup-based tools to improve discoverability and interoperability on websites using the proposed schemas. The rest of this paper is organized as follows. In section 2, we introduce Bioschemas including currently participating organizations as well as some short term planned outcomes. On section 3, we conclude with some final remarks.

2 Bioschemas

Bioschemas is a community initiative aiming to improve data discoverability and interoperability in LS. In order to do so, Bioschemas will encourage content providers to use Bioschema markup to consistently expose structured data on their websites. Schema.org includes contributions from multiple domain specific communities and has already been adopted by 30% of all the webpages. Despite the popularity and simplicity of Schema.org, the LS community is lagging behind with in both contribution to and adoption of Schema.org.

One of the reasons behind the lack of adoption of Schema.org in LS seems to be the lack of specialized terms. Although a protein can be labeled as an *itemtype="http://schema.org/CreativeWork"*, it would be more accurate and useful if it could be labeled as *itemtype="http://bioschema.org/Protein"*. Such Extension of Schema.org towards the LS is one of the main objectives of Bioschemas. Such extensions could deal with two very different content types: common generic types such as materials, events and datasets in LS, and specialized types such as phenotype, gene and protein. In order to achieve a comprehensive yet lightweight coverage of LS, Bioschemas has brought together various organizations such as ELIXIR (https://www.elixir-europe.org/), Pistoia Alliance (http://www.pistoiaalliance.org/), Goblet (http://www.mygoblet.org/), biosharing.org (https://biosharing.org/), bioCADDIE (https://biocaddie.org/), BBMRI-ERIC (http://www.bbmri-eric.eu/) and

EMBL-ABR (https://www.embl-abr.org.au/). It also collaborates with the Health and Lifesciences Schema.org group. As a community project, it is open to any organization related to LS to join the effort.

2.1 Specifications

Bioschemas specifications are the main objective of the current working groups, and are intended to facilitate a consistent and coherent adoption of Bioschemas. A specification comprises an introduction, a data model definition, content guidelines, and examples.

- The introduction presents the problem and the goals to focus on.
- The data model corresponds to a schema to be integrated to Schema.org. It also acts as a skeleton in case a data provider decides to take the markup further, to a more specialized and tailored level. The data model also includes controlled vocabularies useful to describe properties, i.e., relationships, between content types defined in the schema. Suggestions regarding the cardinality of the properties will be provided as well. As much as possible existing vocabularies will be used; however, new ones could be introduced if needed.
- The content guidelines comprise the minimum set of content types and properties to be used by a data provider as well as some additional recommendations.
- Examples will make it easier for data providers to learn how to use the schema on their websites. They will pertain to real use case scenarios.

Schema.org provides a simple lightweight data model; thus, the implementation on a website might differ from data provider to data provider, even within the same domain. Such simplicity is at the same time a strength and a weakness. With the provision of specifications, Bioschemas looks to maximize coherence and consistency across different but related websites.

2.2 A discovery platform

Defining LS common and specialized types is the first step, but further work will be required in order to achieve large scale adoption of Bioschemas. Bioschemas envisions the creation of a system where data providers query a registry in order to retrieve information relevant to a particular entity corresponding to a particular data type, for instance a protein –data type, identified as A4_HUMAN or P05067 –entity. The registry would then identify the corresponding type and swipe all the registered data providers for that data type. The registry response would comprise a collation of all the structured data retrieved from different websites in relation to the initial query. The actors and interactions with each other are depicted in Fig. 1. By providing such a registry, Bioschemas aims to facilitate discovery and validation of Bioschemas compliant websites while contributing to data integration and interoperability.

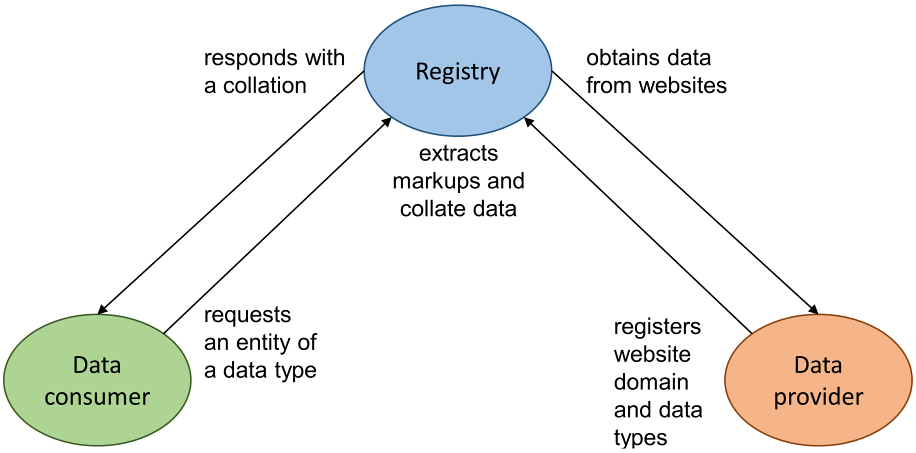

Fig. 1. Bioschemas discovery platform, actors and actions.

Other specialized LS registries can also benefit from this platform as they would become a special type of data consumer. In LS, specialized registries commonly rely on manual processes to update their records. Having a discovery platform would make it make it easier for those registries to collect data that can be later indexed. It would also facilitate the adoption of common metadata specifications as data would be provided following the markup proposed by Bioschemas. The collation made at the registry would take care of basic data cleaning in order to, for instance, avoid duplication.

Third-party applications could also benefit from the Bioschemas discovery platform. For instance, the ELIXIR's Training Portal (TeSS). TeSS provides users with information regarding educational materials and training events. Such data types are covered by the common generic types in Bioschemas. Although it is possible for TeSS to collect such data directly from LS training websites, using the registry would make it easier. Visual components, such as those supported by the BioJS community could also benefit from Bioschemas. BioJS [4] is a community project proving guidelines and tutorials making it easier to create biological data web-based visualization components. Adding Bioschemas markups to BioJS components would be a step forward to make it possible for websites to automatically inject BioJS components corresponding to particular data type.

2.3 Harnessing search engines and their tool ecosystem

Many applications from Google, Microsoft, Pinterest, Yahoo, Yandex and others use Schema.org to power rich, extensible experiences [5]; indeed Schema.org is sponsored by these major search engine providers and operated as an open community. By working closely with a widely used mechanism Bioschemas leverages general infrastructure developed to support structuring data on the Web: standard micro-formats, content management systems, harvesters, validators and search engines. By marking up life science resources with Bioschemas site owners can improve how their sites appear in Google or Bing search results. We have the exciting prospect of using powerful search engines, in native form or especially configured, to perform a rich "search and snippet" service for Life Sciences.

3 Final remarks

Here we have presented Bioschemas, a community effort to extend Schema.org into the LS domain. The success of Schema.org lies in its simplicity and the support provided by major search engines. By extending Schema.org, Bioschemas aims to provide a lightweight semantic layer for LS websites and thus facilitate discoverability and interoperability across them. Visibility is one of the immediate benefits from adopting Bioschemas, as major search engines use the structured data extracted from websites.

Since common and well established ontologies are considered for incorporation in the specifications, Bioschemas should facilitate semantic cross-referencing. Further integration across websites will become possible through a coherent and consistent adoption of the proposed standards. This is particularly relevant for small websites which do not have the capability or resource to provide programmatic access to their data. As more data providers adopt Bioschemas, the cumulative benefits for data discovery and interoperability will become increasingly apparent.

4 References

1. Consortium, T.U., *UniProt: a hub for protein information.* Nucleic Acids Research, 2015. **43**(D1): p. D204-D212.
2. Bento, A.P., et al., *The ChEMBL bioactivity database: an update.* Nucleic Acids Research, 2014. **42**(D1): p. D1083-D1090.
3. Guha, R.V., D. Brickley, and S. Macbeth, *Schema.org: evolution of structured data on the web.* Commun. ACM, 2016. **59**(2): p. 44-51.
4. Corpas, M., et al. *BioJS: an open source standard for biological visualisation - its status in 2014*. F1000Research, 2014. **3**, 55 DOI: 10.12688/f1000research.3-55.v1.
5. Mika, P., *On Schema.org and Why It Matters for the Web.* IEEE Internet Computing, 2015. **19**(4): p. 52-55.