Using community events to increase quality and adoption of standards: the case of Bioschemas

Giuseppe Profiti[1,2,3], Rafael C. Jimenez[4], Federico Zambelli[1,3,5], Ivan Mičetić[1,6], Vito Flavio Licciulli[1,7], Matteo Chiara[1,5], Silvio Tosatto[1,6], Rita Casadio[1,2], Graziano Pesole[1,3,8]

1 ELIXIR-IIB, Italy
2 University of Bologna, Bologna, 40126, Italy
3 Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari, 70126, Italy
4 ELIXIR Hub, Cambridge, CB10 1SD, UK
5 University of Milan, Milan, 20122, Italy
6 Dept. of Biomedical Sciences, University of Padua, Padua, 35100, Italy
7 Institute of Biomedical Technologies, National Research Council, Bari, 70126, Italy
8 Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari “A. Moro”, 70126 Bari, Italy

Abstract

We present how a workshop for the local Italian ELIXIR community steered an improvement of the quality and adoption of Bioschemas, a series of semantic annotation templates for tools, data and samples developed by the ELIXIR Interoperability platform.

Gathering a small number of different end-users and having them focus on applying Bioschemas specification to their tools and data resulted in recruitment of early adopters, eight annotated resources, Bioschemas examples for future users and more than ten suggestions for specification improvement.

This approach could be applied to other open specifications, promoting a wider adoption and the integration of suggestions in a bottom-up fashion.

Keywords

Bioschemas, FAIR, metadata, life science, communities, ELIXIR

Introduction

ELIXIR is a European infrastructure that, among other things, fosters the adoption of common technical solutions and standards in Life Sciences. ELIXIR is managed by a central Hub, with national nodes participating in governance and activities.

ELIXIR is built on five platforms: Compute, Data, Interoperability, Tools and Training. Each platform works on improving a specific aspect of the infrastructure, collaborating with each other to build a common and solid footing for Life Science research in Europe.

The Interoperability platform focuses on standards and technologies for integrating different data sources, tools and building an ecosystem for information exchange and technical solutions.

Bioschemas

Bioschemas (www.bioschemas.org) is a community project supported by the ELIXIR Interoperability platform. It aims to define a set of specifications to improve the semantic annotation and findability of tools, databases, datasets, samples and other elements of a life science workflow (Seibel 2006; Gray 2017) Bioschemas extends and builds on schema.org, an international effort of semantically annotate content on the World Wide Web. Schema.org annotations are widely used by organizations to help search engines in discovering products and services and provide customized views for specific content.

Thus, while objects like books, software and recipes are already covered by schema.org, life science entities like proteins and biological samples are not. Bioschemas' goal is to provide guidelines to help resources in the life science to adopt schema.org to describe minimum information useful for users to find and search data and services. Bioschemas' profiles contribute to provide a minimum and common ground of semantic annotations that help users and integrations resources like search engines and registries to harvest, integrate and present information. Interoperability among Bioschemas is enhanced thanks to a set of standardized and machine-readable properties. Semantic reasoners and other software can also use Bioschemas to automatically infer properties and traverse knowledge thanks to ontological annotations.

Italian workshop

Italy joined ELIXIR in 2016, when some activities were already ongoing. Since then ELIXIR-Italy participated in many ways, from training events, to workshops, to the registration of about 300 tools into the ELIXIR tools registry bio.tools (Ison et al, 2016).

To facilitate the integration of the Italian node into the infrastructure, a series of implementation studies were proposed. One of these implementation studies was aimed to the integration into Bioschemas activities.

The implementation study consisted in two parts: participating in the definition of Bioschemas specifications, and test their implementation feasibility in a subset of ELIXIR-Italy tools and databases. While the main specifications of Bioschemas were reasonably drafted by taking into account the input of big players in the life science community (e.g. PDB, UniProt, tool registries etc), it seemed natural to test how easily they could be used by others, to help the adoption of such technologies also from end-users and small groups.

Results

The workshop produced operational implementations of Bioschemas specifications for several tools and databases. Furthermore, the discussion during the workshop nurtured contributions to Bioschemas community including feedbacks on available documentation, clarifications, fixes and examples, all made possible by the use of an open project infrastructure. These outcomes have been made publicly available through the Bioschemas GitHub repository (https://github.com/BioSchemas).

In particular, the work of the participants led to the following results:

- 6 tools annotations, 5 of which already implemented and used in their respective websites
- 2 databases / data repositories annotations, all already online in their respective homepages
- 1 DataSet annotation, already online
- Contributions to Bioschemas repository
    - 5 tools examples
    - 2 DataCatalog examples
    - a listing of implementing services and their URLs
    - 6 specification proposals
    - 2 identified issues
    - 4 requests for comments or clarifications

More details are available in Table 1.

Submitted proposals included the following:

- suggestions for documentation enhancement relative to controlled vocabularies
- identifiers encoding in DataCatalog profile
- points where the documentations could be clearer (e.g. improved descriptions)

Documentation issues have also been uncovered: as an example, there was a mistake in the specification for the tools, where one key property (url) was replaced by another, unused, property (subjectOf). That specific problem, along with others, has already been tackled by the Bioschemas team.

In general, an overhaul of the documentation was suggested: detailed descriptions of the properties specific to Bioschemas, as opposed to the general schema.org description or very short Bioschemas comments and examples will greatly help developers in adopting the specifications.

Conclusion

The workshop demonstrated how end-users can participate to the definition of specifications and help in the improvement of project documentation. In this case, existing specifications were used by the workshop participants, most of them without any prior knowledge, with minimal effort in their adoption.

The existence of Bioschemas implementations for tools and databases may lift the burden of annotation updates in different repositories, thanks to the ability of crawlers to automatically read user specified annotations (metadata) and update relevant entries in bio.tools or similar registries.

Multiple users implementing the specifications could spot new requirements not only for Bioschemas or the specific addressed projects, but also for basic Bioschemas components like ontologies.

Author contributions

GP organized, chaired the workshop and wrote this paper. RJ co-chaired the workshop and sorted the submitted issues. GP, FZ, IM, VFL and MC implemented Bioschemas in resources from their home institutions. ST is Deputy Head of Node of ELIXIR Italy and suggested the idea of a workshop. RC is the Italian scientific delegate in ELIXIR Board and provided the facilities for hosting the workshop. GrP is the Head of Node of ELIXIR Italy and provided financial support to the event. All authors reviewed the paper and suggested improvements to it.

Acknowledgments

The workshop described in this paper was funded by ELIXIR, thanks to the funding for integrating new nodes into the infrastructure. GP would like to thank the International Master's Degree in Bioinformatics of the University of Bologna for providing the training room.

References

- Seibel, Philipp N., Jan Krüger, Sven Hartmeier, Knut Schwarzer, Kai Löwenthal, Henning Mersch, Thomas Dandekar, and Robert Giegerich. "XML Schemas for Common Bioinformatic Data Types and Their Application in Workflow Systems." *BMC Bioinformatics* 7 (November 6, 2006): 490. https://doi.org/10.1186/1471-2105-7-490.

- Gray, Alasdair JG, Carole Goble, and Rafael C Jimenez. *Bioschemas: From Potato Salad to Protein Annotation*. ISWC 2017 Poster Proceedings, 2017.

- Ison, Jon, Kristoffer Rapacki, Hervé Ménager, Matúš Kalaš, Emil Rydza, Piotr Chmura, Christian Anthon, et al. "Tools and Data Services Registry: A Community Effort to Document Bioinformatics Resources." *Nucleic Acids Research* 44, no. D1 (January 4, 2016): D38–47. https://doi.org/10.1093/nar/gkv1116.

Figures and tables

Result	**Details**	**Notes**
8 annotated resources	• **AGAME** (Tool) • **BAR 3.0** (Tool) • **Cscan** (Tool) • **ITSoneDB** (DataCatalog, DataSet)	Resources in bold already publish Bioschemas annotation online in their respective webpages.

	• **MobiDB** (DataCatalog) • **Pscan** (Tool) • **PscanChIP** (Tool) • SNPs & GO (Tool)	
5 Tools examples	• BAR 3.0 • Cscan • Pscan • PscanChIP • SNPs & GO	https://github.com/BioSchemas/specifications/tree/master/Tool/examples
2 DataCatalog examples	• ITSoneDB • MobiDB	https://github.com/BioSchemas/specifications/tree/master/DataCatalog/examples
“Repository” of services implementing Bioschemas	One CSV file listing services and their URLs	https://github.com/BioSchemas/specifications/blob/master/implementations.csv
6 proposals	• https://github.com/BioSchemas/specifications/issues/117 • https://github.com/BioSchemas/specifications/issues/118 • https://github.com/BioSchemas/specifications/issues/123 • https://github.com/BioSchemas/specifications/issues/126 • https://github.com/BioSchemas/specifications/issues/127 • https://github.com/BioSchemas/specifications/issues/130	
2 issue reports	• https://github.com/BioSchemas/specifications/issues/119 • https://github.com/BioSchemas/specifications/issues/129 *Both are already solved at the time of writing.*	
4 requests for comments or clarifications	• https://github.com/BioSchemas/specifications/issues/124 • https://github.com/BioSchemas/specifications/issues/125 • https://github.com/BioSchemas/specifications/issues/128 • https://github.com/BioSchemas/specifications/issues/131	

Table 1. Workshop results