1st Conference on Research Data Infrastructure

Harmonizing RDM

https://doi.org/10.52825/CoRDI.v1i.280

Published: 07 Sept. 2023

Schema.org as a Lightweight Harmonization Approach for NFDI

Leyla Jael Castro[1][https://orcid.org/0000-0003-3986-0510], Juliane Fluck[1,2][https://orcid.org/0000-0003-1379-7023], Daniel Arend[3] [https://orcid.org/0000-0002-2455-5938], Matthias Lange[3][0000-0002-4316-078X], Daniel Martini[4][https://orcid.org/0000-0002-6953-4524], Steffen Neumann[5][https://orcid.org/0000-0002-7899-7192], Sonja Schimmler[6][https://orcid.org/0000-0002-8786-7250] and Dietrich Rebholz-Schuhmann[1,7][https://orcid.org/0000-0002-1018-0370]

[1] ZB MED Information Centre for Life Sciences, Cologne, Germany

[2] University of Bonn, Bonn, Germany

[3] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

[4] Kuratorium für Technik und Bauwesen in der Landwirtschaft e.V. (KTBL)

[5] Leibniz Institute of Plant Biochemistry, Halle, Germany

[6] Fraunhofer Institute for Open Communication Systems (FOKUS), Berlin, Germany

[7] University of Cologne, Cologne, Germany

Abstract. Schema.org is a controlled vocabulary that makes it easier for web pages to describe their actual content in a semantic, structured and machine-processable way. It is recognized by major search engines and data aggregators, making it easier for researchers to expose metadata describing their research outcomes. Here we present how Schema.org is used (or planned to be used) by some NFDI consortia, becoming a lightweight approach to harmonize digital objects coming from different sources so they can be connected to each other in a meaningful way.

Keywords: Metadata Harmonization, Lightweight Semantics, Schema.org, Metadata, Metadata Schema, Bioschemas

1. Background

Schema.org (from now on SchemaOrg) [1] is a vocabulary collaboratively developed by a community involving major search engines, including Google, Microsoft, Yahoo and Yandex. It offers a simple way for web pages to include structured data markup and thus semantically describe their content. Search engines can use that markup to present results tailored to the nature of the content, and offer added value to end-users. For instance, images are commonly displayed when looking for a recipe so users can get a graphic depiction; related recipes, e.g., including similar ingredients, can also be suggested. Structured markup also makes it possible to create summaries, like the ones displayed when looking for a movie which include similar movies, actors, release year, genre and more.

In recent years, the scientific community, with its ever increasing production of data, has shown interest in SchemaOrg as it presents low adoption barriers to publish data on the web [2]. While development of specialized APIs and web services requires software engineering skills, exposing SchemaOrg structured markup on web pages requires only basic understanding of HTML. As SchemaOrg is compatible with W3C RDF and Linked Data

specifications, the data described with it can be serialized using, for example, JSON-LD (current recommendation by SchemaOrg), and integrated into knowledge-graph-based infrastructures. Additionally, SchemaOrg types and properties can be reused within other RDF-based vocabularies. Another incentive for researchers to use SchemaOrg comes from the Google Data Search [3], a specialized portal released in 2020 helping researchers to find data on the web.

Selecting types and properties best suited to describe scientific outcomes is a different matter and will require some expertise on controlled vocabularies and semantics. Bioschemas [4] is a community project built on top of SchemaOrg, aiming to improve findability of resources in Life Sciences by embedding structured markup on relevant web pages. Bioschemas offers types tailored to Life Sciences but also profiles, i.e., usage recommendations including examples, on top of SchemaOrg types useful to describe scientific outcomes such as datasets, training materials, software and workflows. Other communities such as Science on Schema [5] and the Research Data Alliance Working Group Research Metadata Schemas [6], target particularly datasets and data catalogs.

2. Use of SchemaOrg in NFDI

Multiple NFDI consortia have turned to SchemaOrg as a lightweight approach to harmonize data from the different participant partners. As a side effect, they are also becoming connected along NFDI. Although at a basic level, SchemaOrg markup also contributes to make digital objects findable (via search engines or data aggregators and registries), accessible (exposed over TCP/IP protocol), interoperable (common types, properties and connections to each other), and reusable (via, e.g., inclusion of license and conditions of access).

Most consortia related to Life Science have turned to Bioschemas profiles as they already provide some guidance on how to use SchemaOrg in this domain. For instance, **FAIRagro** will build upon and extend Bioschemas specifications, taking also into account well-known vocabularies in the agri-domain (e.g., AgroVoc [7]). Work on extensions and adoption will involve a variety of domain experts, expert associations and service providers to work collaboratively, via two AgriHackathons. FAIRAgro and DataPlant will also benefit from the work done by the ELIXIR Plant Community [8] which unites a diverse set of services and work on the different implementation studies to increase the Bioschemas compliance of their resources. There are already some first contacts initiated, which will be increased and intensified in the next few years. On its part, **NFDI4Microbiota** is starting with the integration of Bioschemas specifications related to training. **NFDI4Biodiversity** could also benefit from Bioschemas as it offers relevant types such as Taxon and TaxonName. Bioschemas also support chemical related types, MolecularEntity and ChemicalSubstance, that will be useful for **NFDI4Chem** and **NFDI4Cat**. The regular BioHackathons organized by e.g., ELIXIR or the German Node ELIXIR-DE provide opportunities to submit proposals and work on specific needs to improve metadata profiles, data resources or infrastructure.

Outside the Life Sciences domain, **NFDI4Culture** and **NFDI4MatWerk** have been collaborating to create a common ontology including elements from SchemaOrg [9]. Such an ontology can be easily extended to cover more domain-oriented terms, which has been done already in NFDI4MatWerk. Other consortia, such as **NFDI4DataScience** and **NFDI4Memory** are planning to pick up the approach. **NFDI4DataScience** is using SchemaOrg as the default representation for digital objects in their search engine and portal, including training datasets, artificial intelligence models (direct contribution as this object is not yet covered by SchemaOrg core), training and optimization software, and scholarly publications.

3. Future Work

SchemaOrg offers a broad spectrum of types and properties, some of them useful to represent research outcomes, some others that can be combined with domain-specific vocabularies and datasets. This ample coverage in SchemaOrg makes it difficult to use it in a consistent and coherent way (e.g., while someone can use free-text keywords, someone else could favor terms defined in a controlled vocabulary). Bioschemas profiles address this challenge by providing usage recommendations. Finding a way to harmonize across different NFDIs and avoiding duplication of efforts wrt new types and properties could become part of one of the projects in Base4NFDI. Broader adoption will require international acceptance. In the European context, ELIXIR Europe is one of the pioneers using SchemaOrg for science since 2017 [3, 10, 11, 12], while, at an international level, the Research Data Alliance has also contributed in this regard [6].

Data availability statement

No data was used to support the text presented in this abstract.

Author contributions

LJC: conceptualization, project administration, writing – original draft, writing – review & editing. JF, DA, ML, DM, STN, SS: writing - review & editing. DRS: conceptualization, funding acquisition, project administration, writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

NFDI consortia are funded by the Deutsche Forschungsgemeinschaft DFG: NFDI4DataScience project no. 460234259, FAIRAgro project no. 501899475, NFDI4Chem project no. 441958208.

References

1. Guha RV, Brickley D, Macbeth S (2016) Schema.org. Communications of the ACM 59 (2): 44- 51. https://doi.org/10.1145/2844544
2. Gray A, Castro LJ, Juty N, Goble C. (2023) Schema.org for Scientific Data. Artificial Intelligence for Science. World Scientific; pp. 495–514. https://doi.org/10.1142/9789811265679_0027
3. Benjelloun O, Chen S, Noy N. Google Dataset Search by the Numbers. 2020. https://doi.org/10.48550/arXiv.2006.06894
4. Gray AJG, Goble C, Jimenez RC (2017) From Potato Salad to Protein Annotation. ISWC Posters and Demo session. URL: http://ceur-ws.org/Vol-1963/paper579.pdf
5. Shepherd A et al. (2022). Science-on-Schema.org v1.3.0. Zenodo. https://doi.org/10.5281/zenodo.6502539
6. Wu M, Juty N, RDA Research Metadata Schemas WG, Collins J, Duerr R, Ridsdale C, et al. (2022). Guidelines for publishing structured metadata on the Web. RDA. https://doi.org/10.15497/RDA00066
7. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., & Keizer, J. (2013). The AGROVOC Linked Dataset. Semantic Web, 4(3), 341–348. https://doi.org/10.3233/SW-130106

8. Pommier C, Gruden K, Junker A et al. (2021). ELIXIR Plant sciences 2020-2023 Roadmap. F1000Research 2021, 10(ELIXIR):145 https://doi.org/10.7490/f1000research.1118482.1
9. Tietz T, Bruns S, Sack H, Posthumus E. (version 1.1) NFDI4Culture Ontology. Available at https://nfdi4culture.de/ontology
10. García, L. J., Giraldo, O. L., Castro, A. G., & Dumontier, M. (2017). Bioschemas: schema. org for the Life Sciences. In SWAT4LS. https://ceur-ws.org/Vol-2042/paper33.pdf
11. Michel, F. (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
12. Castro LJ, Palagi PM, Beard N, Attwood TK, Brazas MD. (2022) Bioschemas Training Profiles: A set of specifications for standardizing training information to facilitate the discovery of training programs and resources. bioRxiv. https://doi.org/10.1101/2022.11.24.516513