BioHackrXiv

Enabling profile updates through the Data Discovery Engine (DDE)

Ginger Tsueng[1], Sahar Frika[2], Alasdair J. G. Gray[3], Marcos Casado Barbero[4], Alban Gaignard[2], Ivan Mičetić[5], Leyla Jael Castro[6], and Nick Juty[7]

1 The Scripps Research institute **2** Centre national de la recherche scientifique **3** Heriot-Watt University/TPXImpact **4** EMBL-EBI **5** University of Padua **6** ZB MED Information Centre for Life Sciences **7** University of Machester

BioHackathon series:
BioHackathon Europe 2022
Paris, France, 2022
Project 5

Submitted: 04 Apr 2023

Published by BioHackrXiv.org

Abstract

Bioschemas is a grassroots community effort to improve FAIRness of resources in the Life sciences by defining specific Life Science metadata schemas and exposing that metadata from resources that have adopted it. Now that some initial types have been adopted directly into schema.org, an improved mechanism is required to reignite community engagement and encourage profile development. The current process for creating or updating Bioschemas profiles and types is technical and convoluted which creates accessibility issues that can hamper community participation. As adoption of Bioschemas grows and more of the Life Science community considers contributing specific types and profiles, a more accessible creation/modification process is necessary to avoid a loss in engagement. To address this issue, and to drive further Bioschemas adoption, the community has exploited the Data Discovery Engine (DDE) for profile and type development. DDE provides a schema registry and user-friendly tools for creating and editing schemas. The goal of this project is to update existing Bioschemas community profiles in a targeted and crowd-sourced manner, add new profiles as required, and to ensure the documentation is fit for purpose to enable further Bioschemas contributions, at scale.

Introduction

Bioschemas is an opinionated view on schema.org, targeting the Life Sciences community. This entails defining key data and resource 'properties' within specific communities which make those data more findable on the web. Over the past year, a number of these proposed types have been directly incorporated into schema.org. In an effort to build upon this success, we sought to improve the process of proposing specifications ('types' and 'profiles') for new and existing users of bioschemas. The existing process has been deemed cumbersome, complex and convoluted. Furthermore, even when, for example, a profile is updated, there is an additional step required to display it on the bioschemas website.

The Data Discovery Engine (DDE) Schema Playground is a tool for extending schemas from Schema.org. It provides a graphical user interface (GUI) for tailoring schemas from Schema.org to suit the needs of the user and expresses user-generated schemas in JSON-LD/JSON Schema format. The suitability of utilizing the DDE for creating and maintaining Bioschemas specifications was explored as an offshoot effort of Project 29 at the European Biohackathon 2021 [biohackathon2021_29] during which a number of specifications were generated using scripts to convert YAML files generated from a Bioschemas Validation tool to JSON-LD/JSON Schema and tested within the DDE. Since then, ongoing efforts have been made to improve the functionality of the DDE Schema Playground to support the needs of the Bioschemas community.

The focus of Project 5 at the European Biohackathon, 2022 (Tsueng, Frika, et al., 2022) was to simplify the process by which researchers can create and update profiles, and to deploy those changes to the Bioschemas website. In preparation for this event, documentation in the form of four tutorials (Tsueng, Gray, et al., 2022a) was prepared for creating and updating 'types', and for creating and updating 'profiles', where profiles identify key metadata properties within a particular type, and furthermore define the cardinality (number of occurrences), marginality (mandatory, recommended or optional) and value range (expected 'values' such as 'text', 'url', etc). In addition, several automated processes were started for updating the website and DDE Schema Registry when new or updated types and profiles are created; however, these processes were not linked nor officially adopted by the Bioschemas community. While efforts to integrate the automated processes have been ongoing, time zone differences and limitations in developer time to work on these processes have hampered the integration process.

The hackathon setting provided an ideal testing ground for the tutorials, additionally allowing immediate action upon feedback to improve tutorials. While efforts were made prior to the hackathon to automate parts of the process, the hackathon served an instrumental role in providing face-to-face opportunities for integrating the individual automated parts of the process. This integration would otherwise be difficult to do, requiring extensive testing of the different parts of the process, and approval by the Bioschemas steering council for merging the integrated process into the main Bioschemas repositories.

Objectives

1. Refine tutorials based on user feedback
2. Update a number of Bioschemas profiles using DDE and tutorials
3. Invite hackathon participants to work on new profiles and consider the use of Bioschemas for annotating their resouce
4. Engage with ELIXIR Core Data Resources (ELIXIR, n.d.-a) (CDRs) and Deposition Databases (ELIXIR, n.d.-b) (DDs) to push for Bioschemas adoption

Results

Tutorial and manual process improvements

The specification creation and update processes both have manual steps which must be accessible enough to be performed by members of Bioschemas Working Groups and the greater life sciences community. Those members may not necessarily have the technical skills or background to perform the pre-biohackathon process. With this in mind the tutorials were reviewed by those unfamiliar with the process prior to and during the Biohackathon. While tutorial improvements are expected to continue as part of an ongoing iterative process, Biohackathon participants identified process improvements which will be integrated pending wider community discussion and approval. For example, a pull request template (Casado Barbero, 2022) was created during the Biohackathon, which should improve the overall ease and uniformity of the specification creation and update process.

Bioschemas specification updates

Issues with Bioschemas specifications are tracked using GitHub's Issue tracker (Community, n.d.) including many long-standing profile update recommendations. We identified priority or long-standing issues which could be addressed during the course of the Biohackathon and labeled them for easier follow-up. During the course of the Biohackathon, nine GitHub issues were addressed during the creation of 28 updated draft specification JSON-LD/JSON Schema files. Each of these files would be used to test and improve the robustness of the automated processes, in preparation for their integration. Since the Biohackathon, a total of 31 draft

specification JSON-LD/JSON Schema files have been created, displayed on the Bioschemas website and registered in the DDE Schema Registry.

Automated process integration

Each of the 31 draft specification JSON-LD/JSON Schema files represented an opportunity to test the robustness of the automated website conversion script and DDE update script. During the process of testing the scripts, numerous exceptions and bugs were addressed by the >30 total commits made to the repositories. Both the automated website conversion script and DDE update scripts were officially merged into the master/main branch during the Biohackathon. Lastly, the integration was completed by updating the GitHub action for the DDE update script to follow changes made to the repository by the automated website conversion script/GitHub action.

Community engagement and schema adoption

During the initial introduction and midweek progress reports, many groups expressed interest in working with Bioschemas to ensure FAIRness of their hackathon outputs. As seen in Table 1, we met and began discussions with representatives or individual members of these groups with many discussions continuing well after the Biohackathon.

Table 1: Projects and discussions initiated during the Biohackathon

Project #	Project Title	Topic	Follow-up
15	Infrastructure for Synthetic Health Data	Synthetic Health Data	Mappings drafted
09, 17	Disseminating FAIR Machine Learning Models via BioModels, Metadata schemas supporting Linked Open Science (with a focus on reproducibility)	BioModels and BioImaging/ML	Bioschemas Working Group to be created
22	Plant data exchange and standard interoperability	StudyEvent, LabProtocol, Plant	Further discussions needed, profile tweaking
11, 23	Enhancement and Reusage of Biomedical Knowledge Graph Subsets, Publishing and Consuming Schema.org DataFeeds	ChemicalSubstances, MolecularEntities	Profiles updated
32	Training booster: developing FAIR training materials and Learning Paths	BioSamples	Meetings Scheduled; 1st meeting (2022/2/01)

Project #	Project Title	Topic	Follow-up
19	Nightingale 4.0 - Reusable web components for accelerating end-users access to tools platform metadata	interpro applications	Possible profile updates which can be implemented

Discussion

Since the direct adoption, earlier this year, of some existing Bioschemas types into Schema.org, it has become apparent that the process of developing more types needs to be addressed; the pre-existing process was technically and practically convoluted, and had a technical barrier that was unduly high for envisaged users. While this was the primary motivation for this project at the biohackathon, we additionally targeted other activities and tasks to facilitate community engagement and address concerns.

Over the course of the Biohackathon, we have:

1. Improved documentation for existing tutorials, tested in situ by a naive user. Feedback has now largely been incorporated, and updated tutorials are linked here (Tsueng, Gray, et al., 2022b).
2. The process to populate new or updated types and profiles from DDE to the Bioschemas website has been implemented, productionised and tested for robustness.
3. Using (1) and (2), 31 Bioschemas profiles have been updated and exposed through the website.
4. Tackled nine prioritized github issues to do with profiles and types, using DDE.
5. Engaged with a number of resource representatives (Table 1) to incorporate Bioschemas markup, or improve it to updated versions, as appropriate.

The Biohackathon provided a unique opportunity to engage with, in an interactive manner, both a normally distributed implementation team of software engineers, as well as community members. This massively accelerated the speed with which we have engineered process improvements, as well as identifying further emergent communities, and engaging with existing communities to exchange knowledge and needs.

Future directions

While the process for creating and updating Bioschemas profiles has been improved with automation during the Biohackathon, this process currently is only partially automated for the creation and update of Bioschemas types. In the future, the process for creating and updating both Bioschemas types will be included. Further, the current tutorials do not delve into how to use the DDE Schema Playground's validation editor as this tools are subject to improvement.

Code, repositories, and links

- Bioschemas website repository: https://github.com/BioSchemas/bioschemas.github.io
- Bioschemas draft tutorials: https://alasdairgray.github.io/bioschemas.github.io/tutorials/dde/
- Bioschemas specification repository: https://github.com/BioSchemas/specifications
- Bioschemas DDE integration repository: https://github.com/BioSchemas/bioschemas-dde

- Bioschemas project 5 repository for Biohackathon: https://github.com/elixir-europe/biohackathon-projects-2022/tree/main/5
- Bioschemas tutorial feedback: https://github.com/elixir-europe/biohackathon-projects-2022/tree/main/5/feedback
- Project introduction slides: https://docs.google.com/presentation/d/1IJJUtIMR-mE8Zza4Luq-xJtf
edit#slide=id.p1
- Project mid-report slides: https://docs.google.com/presentation/d/1Kh6CscFFqbkOsmRZYkOuaTe
MBijpX-y31colrlfA/edit#slide=id.p6
- Project final report slides: https://docs.google.com/presentation/d/1AuLhyI1V1QNRJFl3jn3uxZSo
edit#slide=id.g188d9bb6b51_0_180
- Project BioHackrXiv paper repository: https://github.com/gtsueng/biohackathon_project5_report/

Contributions

NJ introduced the project and furnished all required progress reports. NJ and GT wrote the BioHackRxiv report. GT, NJ, and AJGG generated updated specifications (profiles and types). GT, SF, and AJGG worked on the integration of the automated processes. NJ, LJG, AJGG, and MCB provided feedback on the tutorials, overall process, and created the pull request template. LJG, GT, and AJGG improved and integrated the tutorials. NJ, GT, AJGG, LJG, IM, and AG engaged in Bioschemas discussions with other projects.

Acknowledgments

Much of this work and key discussions were initiated at the ELIXIR Biohackathon Europe, 2022 held in November. We thank ELIXIR, the research infrastructure for life-science data, for organizing and sponsoring this event which gathered individuals from different communities of practice enabling us to make progress on several Bioschemas community efforts. We thank Egon Willighagen, Sara EL-Gebali, Núria Queralt Rosinach, Marco Brandizi, Rahuman Sheriff Malik Sheriff, Beatriz Serrano-Solano, Nils Hoffmann, Gustavo A. Salazar, Cyril Pommier and Steffan Neumann for engaging in and/or organizing fruitful discussions. Lastly, we would especially like to thank Dana Cernoskova, Katharina Heil, and other members of the Biohackathon organizing committee for their ongoing support throughout the event.

References

Casado Barbero, M. (2022). https://github.com/BioSchemas/specifications/pull/602

Community, B. (n.d.). https://github.com/BioSchemas/specifications/issues

ELIXIR. (n.d.-a). *ELIXIR core data resources*. https://elixir-europe.org/platforms/data/core-data-resources

ELIXIR. (n.d.-b). *ELIXIR deposition databases*. https://elixir-europe.org/platforms/data/elixir-deposition-databases

Tsueng, G., Frika, S., Gray, A. G. J., Casado Barbero, M., Gaignard, A., Mičetić, I., Castro, L. J., & Juty, N. (2022). *Project 5: Bioschemas - enabling profile updates through the data discovery engine (DDE)*. https://github.com/elixir-europe/biohackathon-projects-2022/blob/main/5/README.md

Tsueng, G., Gray, A. G. J., Castro, L. J., & Juty, N. (2022a). *Bioschemas data discovery engine tutorials*. https://alasdairgray.github.io/bioschemas.github.io/tutorials/dde

Tsueng, G., Gray, A. G. J., Castro, L. J., & Juty, N. (2022b). *Bioschemas data discovery engine tutorials*. https://bioschemas.org/tutorials/dde/