Facilitating life science metadata curation through Bioschemas Validators

Deadline for proposals: 1 April 2021

Submission: EasyChair

Review criteria

- A project description, no longer than 200 words
- Which ELIXIR Platforms/Communities the project aligns to
- Who the leads are, including one person who is nominated as primary lead and main point of contact for communications
- If you planning to attend in person, if the project be run virtually. If all project leads cannot attend in person, the person they nominate to help facilitate a hybrid project at the venue
- Expected outcomes
- Number of people/Nodes that will be expected to participate
- Any people (two maximum) from Europe you would like to invite who are critical to the success of the project (name /institute/email)
- Expected outcome and timeframe
- The repository where the code will reside
- How your project would advance if not selected
- How you will advance after the BioHackathon
- If you are planning to attend in person, and if your project could be run virtually

---

Title: Facilitating life science metadata curation through Bioschemas Validators

Abstract *200 words*

Bioschemas profiles are community agreed standards leveraging Schema.org for Life Sciences. They specify the minimal, recommended, optional metadata as well as cardinality and expected reuse of controlled vocabulary. Conformance to these profiles are vital to support harvesting by initiatives such as OpenAIRE.

However, biologists and bioinformaticians may find annotating their resources to be too technically complex and time-consuming without the availability of user-friendly tools.

Multiple initiatives are emerging to provide support tools. FAIR-checker is a web-application, supported by Knowledge Graphs, aimed at providing developers with technical hints to better implement FAIR principles, and provide minimal Bioschemas markup for better findability.

Within the Bioschemas Community, there have been efforts to develop a reusable scraper (BMUSE) to reliably retrieve embedded markup in websites, as well as several validation frameworks to test the conformance of retrieved markup against a stated Bioschemas Profile. These include the TeSS Validator, CTSA/NIH Data Discovery Engine, the ELIXIR JSON Schema validator, and Bioschemas Validata. These frameworks have tried a variety of underlying technologies, including JSON-Schema, ShEx, and SHACL.

The goal of this is project is to leverage Bioschemas community profiles and gather community efforts on metadata validation to provide: scraping and validation tools, basic statistics on live deploys metadata quality (per profile), tools to help the crowd-sourced Bioschemas markup curation.

Keywords: Bioschemas, Knowledge Graph, ShEx, SHACL, SPARQL, JSON Schema

Lead for project.* *name email address (should be corresponding author)*

- Alban Gaignard alban.gaignard@univ-nantes.fr
- Alasdair J. G. Gray
- Leyla Garcia Jael

Expected outcomes.* *Include outcomes expected from Biohackathon (can be in list form) and in what timeframe*

- Tools consuming (machine readable) Bioschemas profiles and producing validation implementations. This will help to validate Bioschemas Live deployments, and compute basic statistics on metadata quality per profiles.
- *Tools helping the curation of Bioschemas annotated resources: (i) ranking resources based on Bioschemas profiles, (ii) randomly picking some resources with urgent curation needs*
- Incorporating the validation into the data ingestion workflow of OpenAIRE

expected participants needed.* *list of expected participants skill set eg researchers developing workflows*

- Chris Child (TeSS portal)
- Ginger Tsueng (Data Discovery Engine)
- Thomas Rosnet (FAIR-checker)
- Alan Williams (WorkflowHub)
- Alessia Bardi (openAIRE)
- Claudio Atzori (openAIRE)
- Nick Juty (Bioschemas)

Number of expected days hacking for project.* *Indicate number of days (4 maximum)*

4

Project Progression (after and without the biohackathon).* *How will the project advance after the Biohackathon and how would it advance if not selected ?*

Validation tools are key components to enhance metadata quality and to ease life science metadata curation. This event is a great opportunity to gather experts and collaboratively develop or update validation tools. If successful, the project will lead to generic tools to be run on community registries and databases already exposing Bioschemas markup (bio.tools, TeSS, etc.).

Attendance at meeting (In person or virtual?).* *We are hoping that the majority of attendees can attend in person this year. If you are planning to lead the project virtually this year, please explain who or how the project would be organised at the hybrid event.*

Both are possible, depending on the sanitary constraints.