Project proposal: Aligning of schemas between (Bio)schemas, Wikidata and beyond

Participants: John Samuel, Andra Waagmeester

Deadline: April 30th

Information required

- Title and description of the project (500 max word limit)

Wikidata is the public knowledge graph of the Wikimedia Foundation. Intended as a data hub for the sister projects such as Wikipedia for articles and Wikimedia Commons for images. However, as an open and accessible source for knowledge Wikidata extends beyond the Wikimedia sister projects.

Various projects such as Gene Wiki, Wikiproject Covid 19, align knowledge with Wikidata and this knowledge is available as linked data for the life sciences. Various tools and methods exist for a straightforward integration of knowledge. One of the challenges remains the alignment of underlining schemas both in the primary sources and in Wikidata. Since there is no central authority that curates its contents - it can be maintained and controlled by anyone - concepts can be described in various patterns or collections of properties.

For any use case it is necessary to understand how applicable concepts are described in Wikidata. In 2019 Wikidata started to support EntitySchemas. These Entityschemas use the Shape Expression (shex.io) language to formally describe data structures in Wikidata.

However, Shape Expressions are only sufficient if there is a large userbase that are proficient in this language. However, learning shape expressions comes with a steep learning curve.

During the Covid-19 Virtual BioHackathon we deployed various methods for schema/shex extraction from existing data sets. These tools are able to extract schema patterns for Wikidata. Still these tools require understanding ShEx to proficiently use them.

During the biohackathon we would like to work on making these tools for extracting shape expression more user-friendly, by creating GUI’s around them intuitive to biomedical scientists to the extent that domain experts don’t need to learn Shape Expression, to understand how information from their and others domain are structured in Wikidata (and possibly by extension in any linked data resource.

    - Include how the project aligns to the ELIXIR goals around platforms and communities
    - The contact details of the person leading the hacking project who will further be the main correspondent (we will only contact this main correspondent
        - Expected outcomes from BioHackathon and beyond
    - Expected Audience: type of participant needed to help with the project
    - Number of days for the hack (Note, it is expected that all participants stay for the 4 days even though some projects may not require the full 4 days for completion)
    - Nominated Participant:Please name who you would like to be funded to attend for your project (this could be an expert in the field or one of the task leads. Max of 2 people should be named)
-
-

Selection criteria

The programme organising committee will select the project based on the following criteria

    - Alignment with ELIXIR goals around platforms and communities
    - Application to broader life science communities
    - Achievability of the timeline of delivery
    - Ambition - is your proposal a cutting edge project applicable to general bioinformatics?
-