Bioschemas Face-to-Face Meeting

Bioschemas Community ELIXIR implementation study

8 May 2019,

Meeting Room 9, Hilton Amsterdam Airport Schiphol

GitHub Usernames

(add yours here to be added as a contributor to use the Kanban boards)

- Matt Styles - @stylesm (added)
- Quentin Groom - @qgroom (added)

Motivations

The last Bioschemas community face-to-face meeting was held in October 2017 (Adoption Meeting). At that time, we approved 12 profiles and proposals for 3 new Schema.org types. We then moved into a period of engagement with the wider community, refinements of our proposed extension for Schema.org, and a broadening of the coverage of the profiles. We have held 12 tutorials to encourage the markup of life sciences resources including a week long hacking activity at the Biohackathon 2018.

An outcome of these engagement and development activities has been to change the nature of the proposed extension from a single "super" type which would be refined through the use of community ontologies to a larger extension for Schema.org. We now have proposals for 30 profiles and 17 Schema.org types with associated properties.

There has been extensive discussion amongst the community through the mailing list and GitHub issues. However, it is now time to come together again and reach consensus for the Bioschemas proposal to extend Schema.org for the life sciences.

Expected Outcomes

This one day meeting will:

- Reach agreement on the Bioschemas types and properties
- Agree pathway for updating profiles
- Plans for Bioschemas community paper
- Discuss future directions for Bioschemas community

Note that this event is not an introductory/tutorial meeting about using Bioschemas.

Registration

Please register for the event using this link.

Attendees are encouraged to apply to the Bioschemas Travel Grant to cover their costs for the meeting.

The meeting was attended by 19 members of the community with representatives from 7 ELIXIR nodes and 3 external organisations.

Agenda

Time	Topic
8:30 - 9:00	Arrival with tea/coffee
9:00 - 10:30	Bioschemas today – Alasdair Overview of Proposed types and properties – Alasdair
10:30 - 11:00	Refreshment break: Hilton Meetings Foyer
11:00 - 12:30	Discussion and agreement of types and properties
12:30 - 13:30	Lunch break: The Bowery Restaurant
13:30 - 15:00	Future directions for the Bioschemas Community
15:00 - 15:30	Refreshment break: Hilton Meetings Foyer
15:30 - 16:30	Drawing up an action plan for Bioschemas ● Updating Bioschemas profiles ● Bioschemas community paper
16:30 - 17:00	Wrap up – Alasdair

Resources

Draft types and properties: http://bio.sdo-bioschemas-227516.appspot.com/ (also https://bioschemas.org/schema/)

Notes

Please use this space to keep notes from the meeting and raise any questions.

Types

Training and Events

Tackling the outstanding Github Issues. The plan, once completed, is to review each schema and see if there are more issues to be addressed. Once completed, publish. Need to reposition Event (strip out training specific properties).

Questions

Topics - is 'about' with a definedTerm still the best practice for defining?

- AG: I believe so

Outcomes:

Course/*Course Instance*

- Profiles updated with a PR to come in the next few days
- Raised issues on schema.org GitHub to have two additional properties added
    - mentions in CourseInstance https://github.com/schemaorg/schemaorg/issues/2246
    - skillLevel in Course https://github.com/schemaorg/schemaorg/issues/2245

*Training Materia*l

Needs a major revision

Event

Needs to be repurposed for conference/workshop/meeting style events. Existing profile should be deprecated with a link added to CourseInstance.

Protein

Protein is here used in its widest possible definition, as classes of amino acid based molecules. Amyloid-beta Protein in human (UniProt P05067), eukaryota (e.g. an OrthoDB group) or even a single molecule that one can point to are all of type schema:Protein. A protein can thus be a subclass of another protein, e.g. schema:Protein as a UniProt record can have multiple isoforms inside it which would also be schema:Protein. They can be imagined, synthetic, hypothetical or naturally occurring.

Note: class concept

GitHub: https://github.com/BioSchemas/specifications/issues/298 → Done in appspot

Actions

- Rewrite Protein profile and enrich with an example(s)
    - GitHub: https://github.com/BioSchemas/specifications/issues/303
    - Template: https://docs.google.com/spreadsheets/d/1UbTy5OACa08ZTX9yk35AaVRUS6YBm1jBqwgG-Y41KGc/edit#gid=1261485211 → Done
    - Updated Protein Type -)

ProteinStructure → BioChemStructure

WwPDB and related data including modelled and hypothetical data. As well as 3D data for chemicals, e.g. stereo chemistry images in an analyzable format (not pure images).

GitHub: https://github.com/BioSchemas/specifications/issues/300 → Done in appspot

Actions

- Rename ProteinStructure to BioChemStructure (discussed with Gene/RNA/DNA, happy with that)
    - GitHub: https://github.com/BioSchemas/specifications/issues/301 → Done in appspot
- Rewrite ProteinStructure/BioChemStructure profile for the PDB case (RNA structure could potentially have a different profile)
    - GitHub: https://github.com/BioSchemas/specifications/issues/302
- Fix typo in property description
    - GitHub https://github.com/BioSchemas/specifications/issues/304 → Done in appspot
- Terms to associate : conformation, protein disordered

Outcomes

- Protein type updated
  Available on appspot
- Protein profile updated
  Available on Bioschemas website
- ProteinStructure renamed to BioChemStructure

Need to review for pushing to schema.org
https://github.com/BioSchemas/specifications/issues/318

Genes

https://docs.google.com/spreadsheets/d/15yAvj5m-Ak_hbkSGrBx2fuhTfcQh-gY1mLT1jdOu5a0/edit#gid=1483018794

- Gene Definition
  A discrete unit of inheritance which affects one or more biological traits, typically associated with a genomic locus https://en.wikipedia.org/wiki/Gene. Examples include FOXP2 (Forkhead box protein P2), SCARNA21 (small Cajal body-specific RNA 21), A- (agouti genotype).
  Now reflected in appspot version
- Missing use cases for Gene Existing Uses cases for genes reviewed and approved. Future work: user communities can define more specific use cases.
  Now available on Bioschemas website
- hasHomolog: `definition: a Gene related to this one by descent, whether in the same species through duplication (paralog) or in another species via a common ancestor (ortholog) or horizontal gene transfer (xenolog). Useful, but not required for an initial release. POSTPONED
  Now reflected in appspot version
- hasOrtholog: we agreed there is no need for this property, on the grounds that a user looking specifically for orthologs or paralogs will be able to differentiate them by the species that the gene belongs to. CLOSED
  Now reflected in appspot version
- isExpressedIn: tissue, organ, etc in which activity of this gene has been observed experimentally. For example HTT (huntingtin) **expressed in brain**. POSTPONED
  Definition has been added to the appspot version
- affectsFunction: (Range: Boolean or text?) Need more information before able to write a formal definition - please close unless there is a strong argument to have this property. (Other 'affects' are already inherited from 'BioChemEntity'). CLOSED
  Reflected in appspot version
- presentInCollection: Also unclear as to what the meaning and use of this property might be - please close as above.CLOSED
  Property removed from appspot version
- partofOperon: To define in a future version.POSTPONED
  Reflected in appspot version

Outcomes

- Gene type reviewed, existing properties are sufficient for an initial version https://github.com/BioSchemas/specifications/issues/317

DNA definition

- Reference to a specific DNA molecule (Deoxyribonucleic Acid). Examples: "Oligo(dT)18 Primer SO131 from ThermoFisher", "Cloning vector PSU2719 DNA"

Outcomes

- DNA definition created

RNA definition

- Reference to a specific RNA molecule (Ribonucleic Acid). Examples: “pBABE-Puro, Retroviral Vector”, “Rosellinia necatrix fusarivirus 1 (RnFV1)”

Outcomes

- RNA defined

DataCatalog, Dataset, DataRecord

Outcomes

- DataRecord sits alongside Dataset in the hierarchy.
- DataRecord links to the Dataset
- Generated RDFa representation of DataRecord
  Available on appspot
- Created schema.org example file for DataRecord
  Available on appspot

Sample/Taxon

Notes from: Matt Styles (Uni of Nottingham/BiobankingUK), Simon Jupp (EBI), Carole Goble (Uni of Manchester/Software Sustainability Institute), Quentin Groom (Meise Botanic Garden)

- Considering Taxon v TaxonName
  - Taxon names can change over time, so proposed that taxon becomes taxon name
  - Discussed that eventually people will have pages on websites which describe taxons, and this is what the taxon type aims to describe, which includes taxon name
  - TaxonName type could include more than just the `name` property - e.g. author, some chronology, etc
  - Discussed possibility of TaxonName being created in addition to Taxon
  - Paused discussion to focus on samples
  - Issue added to GitHub https://github.com/BioSchemas/specifications/issues/309
- Sample Type/Profile - broad discussion
  - Problem with Sample is that it is incredibly broad - could cover human tissue samples, plant samples, etc
  - Discussion over balance between broad enough that it can be applied across multiple concrete types and so specific that it couldn’t be implemented by more than a small number of organisations

    - Discussion over whether Sample Profile is acceptable as it is to cover the bases of different types of samples but then different sample communities can create more specific profiles
    - Some example properties to illustrate..
        - Example of ‘collector’ i.e. Person or Organization who collected the sample - concept of ‘Unknown’ for where this data isn’t known or available
        - Example of ‘dateCreated’ or ‘dateCollected’ which should be applicable to all types of samples but may be difficult if the date is simply known as e.g. ‘the 1990s’.
    - [WIP] Current thinking is add additional properties to the Sample Type even if they are only relevant to a small community or group of people; on top of this create [flat structure] additional Sample Profiles e.g. TissueSample, PlantSample.
    - Discussion over the concept of ‘ordering’ e.g. the order of authors in a MANY author property. Agreed that this is up to conventions of those specifying and consuming, rather than something to encode in schema itself.
    - Reiterate that examples and implementations are important.
    - Agreed like the approach of more optional properties and keeping general rather than adding lots of specific recommended or required properties.
- Strawman minimal properties for a Sample
    - Required
        - Identifier
        - name
        - description
    - Recommended
    - Optional
        - collector [the person or persons who created the sample]
        - CreationDate [the date the sample was created]
        - geographicLocation [the location of origin of the sample] (type:Place)
        - taxonomicRange [Text or URL or bioschemas:Taxon]
        - material [https://schema.org/material - A material that something is made from, e.g. heart, stem cell, whole plant, soil…]
        - age [how old the object was when sample was created]
        - phenotype [range Text or URL or schema.org Phenotype]
        - associatedDisease [the disease that is represented by the sample e.g. malaria, bowel cancer, coeliac disease, Creutzfeldt–Jakob disease, etc]
        - gender = https://schema.org/gender
        - pedigree
        - isControl [A boolean (true/false) value. If true, this sample is being use as a normal control, often in combination with another sample]
- GBIF would be a good example
    - They may need some examples to go from
        - Schema.org, e.g. for PostalAddress https://schema.org/PostalAddress
        - Soon to be on https://directory.biobankinguk.org
        - On the Bioschemas GitHub repo, e.g. for Event: https://github.com/BioSchemas/specifications/tree/master/Event/examples
- Whats the examples - from schema.org - assume implementation by variant hacking of example.

- Must be in the bioschema github example and easier to find.

Outcomes

- Sample type renamed to BioSample and inherits from BioChemEntity
  https://github.com/BioSchemas/specifications/issues/306
  Available on appspot
- Suggested properties to be added to the BioSample type and harmonised these with those in the BioChemEntity where necessary
- Generation of community specific profiles to be done, e.g. HumanSample profile
- Examples developed
  https://github.com/BioSchemas/specifications/issues/299
- Propose that taxonomicRange property should be added to Dataset.
  - Dan questioned whether this could be achieved using the existing https://schema.org/about property

Profile Status

Questions from Alasdair

- Can we deprecate existing “current” profiles, particularly those that do not have any (much) uptake?
- Should types be included in schema.org before a profile becomes current/1.0?
- Can we require 2 independent live deploys before a profile becomes current?
- Should we be versioning our URLs, e.g. https://bioschemas.org/specifications/DataCatalog/0.1/?
- Can we move examples into versioned folders?

Community Development Process

- Use Define Bioschemas types and properties project kanban board on GitHub to track development of types and properties. Particularly
  - ‘Backlog’ which contains potential ideas for types and properties
  - ‘For next release’ contains the cards that are to be worked on for the coming release
  - ‘Done for next release’ contains cards once they are completed and ready for release
- Monthly updates
- Quarterly updates to Schema.org
  - Look at using some of the ideas in the TensorFlow RFC to structure pushes to Schema.org

Logistics

Meeting Requirements

We are planning to hold the event at Schiphol Airport. We will have a meeting room for up to 50 people with standard AV equipment for presenting slides. Internet access will be available for attendees. We will need flipcharts/whiteboard to support ad hoc discussions.

Estimated budget

70 EUR (or 60 GBP) per delegate package

50 participants

Estimated cost: 3500 EUR

BioSample Example

The following example has been copied across to the Samples directory on GitHub. https://github.com/BioSchemas/specifications/blob/master/Sample/examples/MeiseBotanicGardenSpecimen-microdata.html

http://www.botanicalcollections.be/specimen/BR0000009119529

```

<div itemscope itemtype="https://bioschemas.org/types/Sample/">

  <span itemprop="taxonomicRange">Abildgaardia disticha Lye</span>

<link itemprop="" />

<div itemscope itemtype="https://bioschemas.org/types/Sample">

   <span itemprop="creator" itemscope itemtype="https://schema.org/Person">

      <span itemprop="name">Harley R.M.</span>

      <span itemprop="identifier">http://purl.oclc.org/net/edu.harvard.huh/guid/uuid/82261503-351e-40c1-bfcb-f21bc248e86d</span>

   </span>

   <span itemprop="taxonomicRange">Shrub</span>

</div>
```

```
<meta itemprop="phenotype" content="flowering" />
        itemtype="http://schema.org/OrganizationRole">
    <div itemprop="member" itemscope
            itemtype="http://schema.org/Person">
      <span itemprop="name">Joe Montana</span>
    </div>
    <span itemprop="startDate">1979</span>
    <span itemprop="endDate">1992</span>
    <span itemprop="roleName">Quarterback</span>
</div>

```