Metadata specification and relations to other models

Susanna-Assunta Sansone, PhD
Philippe Rocca-Serra PhD,
Alejandra Gonzalez-Beltran, PhD

and the Metadata WG members

ELIXIR All Hands Meeting, Barcelona, 10 March, 2016
Synergies with many groups, including:

- BD2K Center for Expanded Data Annotation and Retrieval (CEDAR)
- BD2K cross-centers Metadata WG
- ELIXIR EXCELERATE WP5 Interoperability

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
Define a *metadata specification* that support intended capability of the Data Discovery Index (DataMed) prototype to harvest, e.g.

- key experimental and data descriptors, such as relations between authors, datasets, publication and funding sources, nature of biological signal, nature of perturbation etc.
Define a *metadata specification* that support intended capability of the Data Discovery Index (DataMed) prototype to harvest, e.g.

- key experimental and data descriptors, such as relations between authors, datasets, publication and funding sources, nature of biological signal, nature of perturbation etc.

Use cases and the competency questions used throughout

- To define the appropriate boundaries and level of granularity: which queries will be answered in full, which only partially, and which are out of scope

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
WG3 Metadata - Phase 1, completed

*Metadata specification v1*, future-proofed for progressive extensions, to support intended capability of the DDI prototype

**PHASE 1 OUTPUT:**
  - The WG3-MetadataSpecifications-v1.zip contains a *document*, two *Appendixes*, *JSON schema* and *examples*.

If you wish to provide *comments* on this document, please, use the *live Google version* (no login required). If you are a WG3 member, use the mailing list; if not, please send your comments to biocaddie[at]ucsd.edu.
WG3 Metadata - Phase 1, completed

**Metadata specification v1**, future-proofed for progressive extensions, to support intended capability of the DDI prototype

**PHASE 1 OUTPUT:**

  - The WG3-MetadataSpecifications-v1.zip contains a document, two Appendixes, JSON schema and examples.

If you wish to provide comments on this document, please, use the live Google version (no login required). If you are a WG3 member, use the mailing list; if not, please send your comments to biocaddie[at]ucsd.edu.

**Created using 2 complementary approaches**

- **top-down:** analyzing use cases
  - Competency question
    - Search for *organism* x in *biological process* y (apoptosis) at *scale* z with an estimate of the reliability of the annotations
    - Search for new *drug* x to predict and track *biological process* x (cardiotoxicity)
    - Search for *data type* x (omics correlates) of *biological process* for *drugs related to drug* x
    - Search for *data types* a, b, and c (EHR data, self-report, sensor) to determine natural history of patients given *drugs similar to drug* x
    - Track responses to treatment to ensure detection of *biological process* x
    - Find *patient data* “like these” with similar treatments, responses to treatment, genetics
    - Search for *studies* a-z with *patient data* with *biological process* x (e.g., obesity as measured by BMI) and *interventions* a-z. Then filter on demographic characteristics.

- **bottom-up:** mapping existing standards/schemas

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
Bottom up approach: schemas evaluated

- schema.org
- DataCite
- RIF-CS
- W3C HCLS dataset descriptions
- ISA
- BioProject
- BioSample
- MiNIML
- PRIDE-ml
- MAGE-tab
- GA4GH metadata schema
- SRA xml
- CDISC SDM / element of BRIDGE model

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
These metadata is either too much or too little

- Many databases won’t have all these metadata elements
- Conversely, domain-specific databases (e.g. focusing on a type of study, organism or technology) have more detailed metadata
We already know that one size does not fit all

- These metadata is either too much or too little
  - Many databases won’t have all these metadata elements
  - Conversely, domain-specific databases (e.g. focusing on a type of study, organism or technology) have more detailed metadata
- We need to refine the core and boundaries for the DDI
  - we have aimed to have *maximum* coverage of use cases with *minimal* number of data elements
  - we do foresee that not all questions can be answered in full
Next steps and relation to bioschema.org

- We are finalizing the Metadata specification v1.1
  - Release mid March and open to community comments for 2 weeks via GitHub and Google docs - links from WG3 homepage

- Next steps will be packaging and releasing of v1.2 by the end of April also via
- It will also include definition and examples of the proposed DATaset Tag Suite format (in JSON and/or serializations) for a scalable way to index data sources in the DataMed prototype

- Additional step could be mapping to schema.org to identify ‘missing’ elements and create an extension as part of bioschema.org
Next steps and relation to bioschema.org

- We are finalizing the Metadata specification v1.1
  - Release mid March and open to community comments for 2 weeks via GitHub, Google docs - links from WG3 homepage
- Next steps will be packaging and releasing of v1.2
  - by the end of April also via GitHub, Google docs and Zenodo
  - it will also include definition and examples of the proposed `<DATS>` DATaset Tag Suite format (in JSON and/or serializations) for a scalable way to index data sources in the DataMed prototype
Next steps and relation to (bio)schema.org

- We are finalizing the Metadata specification v1.1
  - Release mid March and open to community comments for 2 weeks via GitHub, Google docs - links from WG3 homepage

- Next steps will be packaging and releasing of v1.2
  - by the end of April also via GitHub, Google docs and Zenodo
  - it will also include definition and examples of the proposed <DATS> DATaset Tag Suite format (in JSON and/or other serializations) for a scalable way to index data sources in the DataMed prototype

- Additional step will be mapping to schema.org
  - to identify ‘missing’ elements and create an extension as part of bioschemas.org