The dataMED DATS model annotated with schema.org

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

Oxford e-Research Centre, University of Oxford, UK

Supported by the NIH grant 1U24 AI117966-01 to UCSD

PI: Lucila Ohno-Machado, Co-Investigators at:

University of California, San Diego
University of Texas Houston
University of Michigan
University of Oxford

EXECUTIVE COMMITTEE

George Alter - ICPSR, University of Michigan
Ian Fore - NIH
Jeffrey Grethe - University of California San Diego
Lucila Ohno-Machado - University of California San Diego
Susanna-Assunta Sansone - University of Oxford and Nature Publishing Group
Hua Xu - University of Texas Houston

STEERING COMMITTEE

George Alter - ICPSR, University of Michigan
Dianne Babski - National Library of Medicine
Ian Fore - NIH
Jeffrey Grethe - University of California San Diego
Jennie Larkin - NIH
Dawei Lin - NIH
Ron Margolis - NIH
Lucila Ohno-Machado - University of California San Diego
Susanna-Assunta Sansone - University of Oxford and Nature Publishing Group
Heldi Sofia - NIH
Hua Xu - University of Texas Houston
Alison Yao - NIH
Like **JATS** (Journal Article Tag Suite) is used by **PubMed** to index literature, **DATS** (DatA Tag Suite) is needed for a **scalable** way to index data sources in the **DataMed** prototype.

A community effort
What is **DATS** support to do and be?

- Enabling **discoverability**: find and access datasets

- Focusing on surfacing **key metadata descriptors**, such as
  - information and relations between authors, datasets, publication, funding sources, nature of biological signal and perturbation etc.

- *Not* the perfect model to represent the experimental details
  - the level of details and metadata needed to ensure interoperability and reusability are left to the indexed databases

- Better than just having keywords
  - we have aimed to have *maximum* coverage of use cases with *minimal* number of data elements and relations
The development process in a nutshell

Metadata elements identified by combining the two complementary approaches

USE CASES: top-down approach

SCHEMAS: bottom-up approach

**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.

**DATS**

DatA Tag Suite

(v1.0, v1.1, v2.0, v2.1)

Model serialized as JSON schemas and mapping to schema.org
Standing on the shoulders of giants

- schema.org
- DataCite
- RIF-CS
- W3C HCLS dataset descriptions (mapping of many models including DCAT, PROV, VOID, Dublin Core)
- Project Open Metadata (used by HealthData.gov is being added in this new iteration)

- ISA
- BioProject
- BioSample

- MiNIML
- PRIDE-ml
- MAGE-tab
- GA4GH metadata schema
- SRA xml
- CDISC SDM / element of BRIDGE model
Convergence of elements extracted from competency questions and existing (generic and biomedical) data models (incl. DataCite, DCAT, schema.org, HCLS dataset, RIF-CS, ISA-Tab, SRA-xml etc.)

DATS DatA Tag Suite model for scalable indexing

core entities

- DataSet
- DatasetDistribution
- DataRepository
- Publication
- DataStandard
- Software
- License
- Dimension
- DataType
- Access
- Grant
- Annotation
- Organization
- Person
- Material

extended entities

- Instrument
- AnatomicalPart
- MolecularEntity
- TaxonomicInformation
- BiologicalEntity
- CategoryValuesPair
- StudyGroup
- Treatment
- Disease
- Activity
- DataAnalysis
- Study
- DataAcquisition

Adoption of elements extracted from and from DataCite and from FORCE11 DATA CITATION IMPLEMENTATION PILOT PROJECT (DCIPP)
General design of the **DATS**

**Dataset**, a core entity catering for any unit of information

- archived experimental datasets, which do not change after deposition to the repository => examples available for dbGAP, GEO, ClinicalTrials.org
- datasets in reference knowledge bases, describing dynamic concepts, such as “genes”, whose definition morphs over time => examples available for UniProt

**Dataset** entity is also linked to other digital research objects

- **Software** and **Data Standard**, which are also part of the NIH Commons, but the focus on other discovery indexes and therefore are not described in detail in this model
Serializations and use of schema.org

- DATS model in JSON schema, serialized as:
  - JSON* format, and
  - JSON-LD** with vocabulary from schema.org
    - serializations in other formats can also be done, as / if needed

- Benefits for DataMed and databases index by DataMed
  - Increased visibility (by both popular search engines), accessibility (via common query interfaces) and possibly improve ranking

- Extending schema.org
  - Submitted to their tracker missing DATS core elements
  - Coordinating via the bioschemas.org initiative (ELIXIR is also part of) the extension of schema.org for life science

---

* JavaScript Object Notation
** JavaScript Object Notation for Linked Data
DATS DatA Tag Suite

core and extended elements

Diagram showing relationships between various entities such as Dataset, License, DataRepository, Software, Disease, BiologicalEntity, AnatomicalPart, TaxonomicInformation, Study, MolecularEntity, Instrument, and more.
Core elements provide the basic info

- **What** is the dataset about?
  - *Material*

- **How** was the dataset produced? *Which* information does it hold?
  - *Dataset / Data Type* with its *Information, Method, Platform, Instrument*

- **Where** can a dataset be found?
  - *Dataset, Distribution, Access objects* (links to *License*)

- **When** was the dataset produced, released etc.?
  - *Dates* to specify the nature of an event {create, modify, start, end...} and its timestamp

- **Who** did the work, funded the research, hosts the resources etc.?
  - *Person, Organization* and their roles, *Grant*
Of the 18 core elements none is mandatory
Only few properties of the 18 core elements are mandatory.
To **evaluate** DATS model capabilities

Other adopters **exporting** DATS in their APIs

**Work in progress:** documentation and curation guidelines for adopters
relations to other BD2K efforts

- For datasets not yet in a formal repositories
  - CEDAR metadata authoring tool can be used to provide DATS-compliant metadata to be later indexed by DataMed

- Mapping DATS to omicsDI model
  - To be able to index datasets in this aggregator

- Ensure that the citation metadata for repositories’ landing page maps to core DATS elements
Interlinking data MED to other indexes
Interlinking [data MED] to other indexes
WORKGROUP 3 GROUP LINKS

LATEST VERSION:

  - Live google document; comments can be added, no login required
  - Specification and Appendixes also available in the bioCADDIE GitHub repository, along with JSON schemata, examples and schema.org annotated JSON-LD serialization (also citable as DOI: 10.5281/zenodo.62024)

Questions to biocaddie[at]ucsd.edu (with subject line: DATS spec).