The DataMed DATS Model, Annotated With Schema.org

Susanna-Assunta Sansone, University of Oxford; Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, George Alter, Mary Vardigan, Jeffrey Grethe, Hua Xu, and Members of the DataMed Development Team and of the bioCADDIE Working Groups

The NIH BD2K bioCADDIE's DataMed stores metadata generic enough to describe any dataset using a model we have called the DAta Tag Suite (DATS). Akin to the Journal Article Tag Suite (JATS) used in PubMed, DATS enables submission of data for "ingestion" by DataMed.

DATS is a community-driven model designed to cover both (i) experimental datasets, which do not change after deposit in a repository, and (ii) datasets in reference knowledge bases describing dynamic concepts, whose definition morphs over time. The DATS model has a core and extended set of elements, to progressively accommodate more specialized data types. Like the JATS, the core elements are generic and applicable to any type of datasets. The extended DATS includes an initial set of elements, some of which are specific for life and biomedical science domains and can be further extended. The DATS entities are available as machine-readable JSON schemata, with examples and schema.org annotated JSON-LD serialization (DOI: 10.5281/zenodo.62024 and https://github.com/biocaddie)

The DATS model was developed given the following considerations:

- A variety of data discovery initiatives exists or are being developed, different scope, use cases and approaches. Their metadata schemas are valuable and were reviewed to determine essential items. The metadata schemas and models used in the mapping have been described in the BioSharing Collection (https://biosharing.org/collection/bioCADDIE)
- Identification of the initial set of metadata elements was based on: (i) analyses of use cases (a top-down approach); and (ii) mapping of existing metadata schemas (a bottom-up approach). From the use cases, a set of 'competency questions' were derived - these defined the questions that we want DataMed to answer- abstracted, key concepts binned in entities, attributes and values categories, to be easily matched with the results of the bottom-up approach.