DATS Specification v2

NIH BD2K DataMed model annotated with schema.org

Status:

Working version 2, released on 1 June 2016.

Scope of the Document:

Section 1 of this document introduces the **DatA Tag Suite** (**DATS**) model, its JSON-schemas-based serialization, a mapping between DATS and schema.org and a schema.org-annotated JSON-LD context and a description of the next steps; section 2 provides an overview of the methods leading to the DATS development, including the use cases and the existing schemas and material reviewed.

Abstract:

The DATS model has been designed to support the intended capability of the NIH BD2K Data Discovery Index prototype, outlined in the bioCADDIE White Paper and named **DataMed.** The DATS model and its serialization describe the metadata and the structure for datasets needed to populate the DataMed prototype, along the line of the Journal Article Tag Suit (JATS) used by PubMed for literature. The DATS has been designed to have core and extended elements, to progressively accommodate more specialized data types, as needed.

Associated Material:

Appendix I, II and III (related to the current DATS model v2), Appendix IV (schemas mapping, part of the development process), along with the schema.org-annotated JSON serialization and examples (of core and extended DATS) are available from the bioCADDIE Github repository. Previous versions, presentations and notes on this work can also be found at the bioCADDIE Metadata Working Group (WG) 3 webpage. The schemas and models in Appendix III are also described in the BioSharing Collection for bioCADDIE.

Intended Audience:

This document is aimed at: (i) the DataMed development team that will implement and test the model, (ii) prospective data sources that wish to be indexed in the DataMed prototype, including those attending the DataMed repository workshop in June 2016, and (iii) developers of data harvesting and other metadata tools and catalogs.

Authors and Contact:

The work has been designed and carried out - as an *open activity* with *community input* - by core bioCADDIE project' members (Susanna-Assunta Sansone, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, with George Alter, Mary Vardigan, Jeffrey Grethe, Hua Xu and other members of the DataMed development team), with contributions from the members of the bioCADDIE Metadata WG 3 and WG7. Comments to the DATS can also be added to this

Google document, to the Github issue tracker or sent to Susanna-Assunta Sansone (WG3 chair) via biocaddie[at]ucsd.edu (subject line: DATS Spec V2).

Table of Content

1. The DATS Model

This section and related Appendixes describe the DATS model, providing a summary of the previous versions, an overview of the key metadata elements and their relations, and a detailed description of each entity. A full description of the DATS core and extended elements is in a separate ***"Appendix I - NIH BD2K BioCADDIE DataMed DATS model v2 file"*** available as a Google spreadsheet, and also linked availbale from the bioCADDIE Github repository. The methods used for its development, including the use cases and the existing schemas and material reviewed, are described in section 2.

1.1. Brief History Leading to this Version

In August 2015, the specification v1.0 (DOI:10.5281/zenodo.28019) was released, with the model available as machine readable JSON schemata and several examples. The initial model was tested by the DataMed development team with a variety of data sources. Following the evaluation phase and review by the WG3 members and the larger community, several changes and additions were made leading to the release of the specification v1.1 (DOI:10.5281/zenodo.53078) in March 2016.

The current v2 represents a further evolution of the model, including the accessibility metadata elements produced by bioCADDIE WG7, a set of JSON-schemas and JSON examples, a schema.org-annotated context file (based on the DATS to schema.org mapping), as a starting point for a JSON-LD serialization. The latter activity, identifying additional elements to extend the schema.org model, is also carried out under the bioschemas.org umbrella.

1.2. General Design

The DATS model is designed around the *Dataset*, an element that intends to cater for any unit of information stored by repositories. *Dataset* covers both (i) experimental datasets, which do not change after deposition to the repository, and (ii) datasets in reference knowledge bases, describing dynamic concepts, such as "genes", whose definition morphs over time. The *Dataset* element is also linked to other digital research objects part of the NIH Commons, such as *Software* and *Data Standard*, which are the focus on other discovery indexes and therefore are not described in detail in this model. The model may appear quite detailed in places as consequence of (i) the combined approaches used to identify the required metadata elements (see section 2), and (ii) the attempt to aim for the maximum coverage of use cases with minimal number of metadata elements. Nevertheless, it is anticipated that not all use cases can be fulfilled, and that it is difficult to foreseen all type of data sources the DataMed prototype should retrieve information from.

1.3. Core and Extended Elements

The DATS model has core and extended elements, to progressively accommodate domain-specific metadata for more specialized data types, as needed. Like the JATS, the core DATS 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, environmental and biomedical science domains and can be further extended as needed. Note that the core should

by no mean seen as the mandatory set; further information on the requirement level is provided in section 1.5. An overview of the core and extended elements, their types and relations are shown in Figure 1 and 2, respectively. The distinction between core and extended DATS will be reviewed also after the DataMed repository workshop in June 2016.

Figure 1. A schematic overview of the **DATS core elements**, their types and relations:

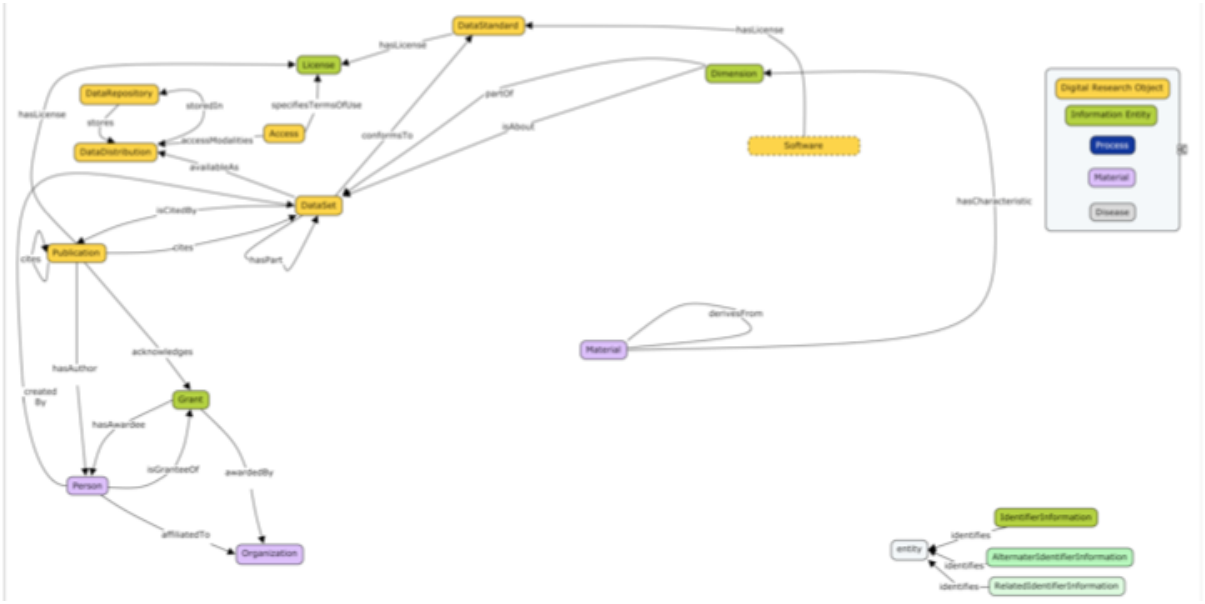
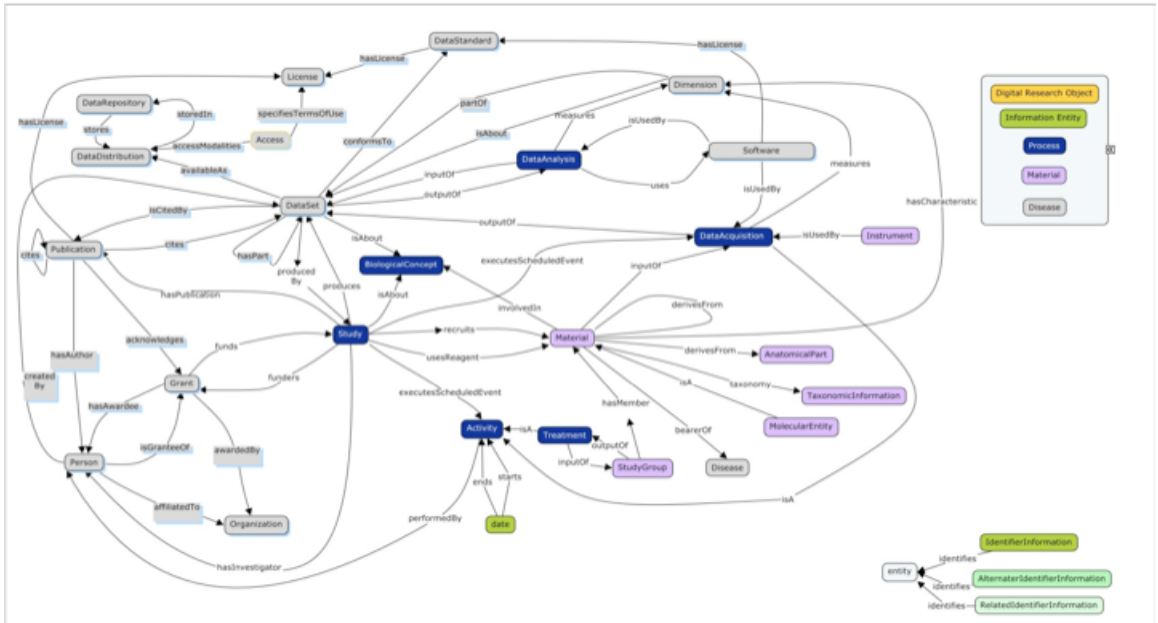


Figure 2. A schematic overview of the **DATS core and extended elements**, their types and relations:



1.4. Towards a Schema.org Annotated JSON-LD Serialization

The core and the extended DATS entities are made available as machine readable JSON schemata, which will be annotated with schema.org elements and made available in the bioCADDIE Github repository, which also includes JSON examples of existing datasets. The use of schema.org annotation - and its extension to the life and biomedical areas - is an ongoing discussion under the bioschemas.org umbrella and NIH Commons Working Groups, which the bioCADDIE team is part of. Preliminary discussions indicate that by implementing a schema.org-annotated DATS model, the DataMed prototype will benefit from an increased visibility (by search engines and tools), increased accessibility (via common query interfaces), and possibly, an improve in search ranking. Data repositories indexed by DataMed will also get the benefit of being more visible to search engines through DataMed. The initial ***"Appendix II - NIH BD2K BioCADDIE DataMed DATS v2 mapped to schema.org file"*** types and properties is available as a Google spreadsheet; this may be subject to change as current schema.org elements and those of health and life science extension evolve, also because the current life science extension is biased towards clinical studies. Gap in coverage have been identified during the DATS annotation process, and are highlighted in the Appendix II file; notifications of these gaps will be submitted to the schema.org github tracker.

1.5. Detailed Model Description

The file describes the metadata elements, grouped by types, and the Google JSON style guide has been used to name relevant elements. The descriptors for each metadata element (Entity), include: Property (describing the Entity), Definition (of each Entity and Property), Value(s) (allowed for each Property).

In Appendix I, the DATS elements are also associated to relevant use cases (detailed in section 2.2) and/or to the existing schema(s)/model(s) used in the development process (detailed in section 2.3), to justify their relevance and provenance.

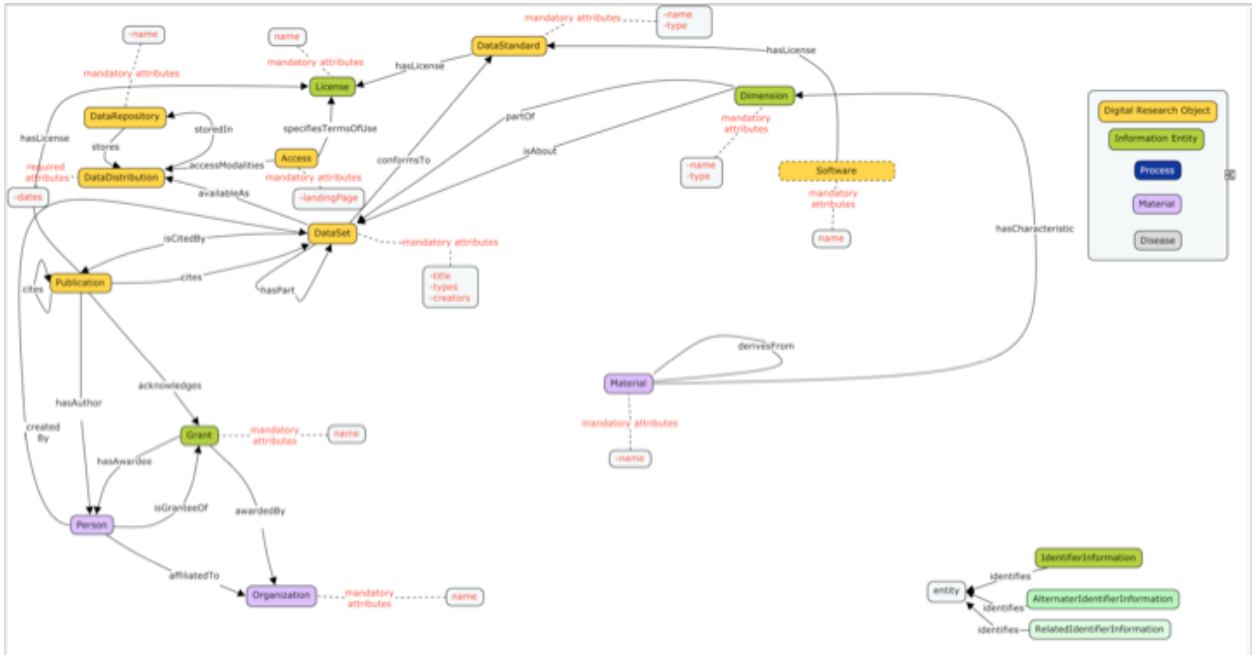
1.5.1. Cardinality and Requirement Level

In both core and extended DATS, Entity are **not** mandatory but applicable only if relevant to the dataset to be described; when an Entity is used, also only few of its Properties are defined as mandatory. Figure 3 provides a view of the core entities highlighting the few properties defined as mandatory.

Cardinality restrictions indicate the number of valid occurrences for an attribute; the initial set of metadata elements are ranked and provisionally associated with a requirement level. The keywords "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119]. The requirement level for a field that is dependent on another one appears in between parenthesis, e.g. 'identifierScheme' MUST be present if 'identifier' is available, and we indicate this with (MUST). While there is some overlap in these specifications (e.g. if the cardinality of an attribute is 1, the requirement level is necessarily MUST), the requirement level adds information about the relative importance of including or not the non-compulsory attributes (either because they are recommended or they

are truly optional). Cardinality restrictions will be used for data modelling purposes. The requirement levels will be reviewed also after the DataMed repository workshop in June 2016.

Figure 3. A schematic overview of the **DATS core entities** and theirs **few properties with requirement level "MUST":**



1.5.2. Mapping DATS Back to Existing Schemas/Models

To facilitate the use of the DATS model by prospective data sources that wish to be indexed in the DataMed prototype, the DATS elements have been mapped to a number of existing (i) generic and widely used schemas, such as schema.org and DataCite, and (ii) domain specific repositories schemas. This mapping is in a separate ***"Appendix III - NIH BD2K BioCADDIE DataMed DATS v2 mapped to other models file"*** available as a Google spreadsheet and also linked availbale from the bioCADDIE Github repository. The work is also being carried out by the DataMed development team to inform the implementation of data harvesting converters, pulling data from those repositories.

2. Standard Operating Procedure

This section outlines the methods and the process used to identify an initial set of metadata elements, leading to the specification v1.0, v1.1 and subsequent current v2.

2.1. Combined Approaches

A variety of data discovery initiatives exists or are being developed; although they have different scope, use cases and approaches, the analysis of their metadata schemas has been a valuable guidance. Several metamodels for representing metadata also exist and have been reviewed. In

addition to these, the results of the following approaches has been compared and combined to identify the initial set of metadata elements:

- an analysis of the use cases (top-down approach; section 2.2); and
- a mapping of existing metadata schemas (bottom-up approach; section 2.3 and Appendix III).

2.1.1. Data Discovery Initiatives and Metadata Initiatives

This is a **non-comprehensive** list of the data discovery and integrative initiatives analysed, which might have more specific aims and different use cases than the intended capability of the DataMed prototype.

1. UK JISC Research Data Registry and Discovery Service: relies on Registry Interchange Format Collections and Services (RIF-CS); related documentations: Github repository, WP3: Metadata Development and Standardisation, Report: metadata mapping schemes / recommendations (version 9, 2014-05-09).
2. Datacite Metadata Search to search datasets registered with Datacite.
3. European EUDAT B2FIND: relies in CKAN (CKAN Dataset Model) and harvest data using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Other references: documentation and mapping files.
4. Research Data Alliance (RDA) Research Data Switchboard relies on OAI-PMH protocol (Github repository).
5. National Data Service.
6. National Institute of Health's Neuroscience Blueprint funded Neuroscience Information Framework (NIF).
7. The National Institute of Diabetes and Digestive and Kidney Diseases' NIDDK Information Network (dkNET).
8. Data Documentation Initiative Draft Specification of DDI-RDF Discovery Vocabulary (Disco) for the discovery of microdata sets and related metadata using RDF technologies in the Web of Linked Data (that relies on the Data Cube, DCAT and XKOS).
9. Global Alliance for Genomic Health: specifically Data Working Group, Genotype-2-phenotype task team, NIH Office of Director funded Monarch Initiative - these are all efforts to develop standardized schemas for genotype-phenotype data integration and data sharing.
10. eTRIKS standards starter pack, under a pre-competitive private, public Innovative Medicine Initiative - aiming to bridge clinical/CDISC, ISA and other community-based standards.
11. RDA Working Group on Data Description Registry Interoperability.
12. EBI RDF Platform.
13. Content Standard for Digital Geospatial Metadata Part 1: Biological Data Profile, 1999.
14. Open PHACTS Discovery Platform - integrating pharmacological data resources according to Dataset descriptions for the Open Pharmacological Space, based on W3C VoID for describing Linked Datasets, HCLS Dataset descriptions by the W3C Semantic Web in Health Care and Life Sciences Interest Group.
15. Just Enough Results Model (JERM) from the SEEK for Science project.

16. Metadata for data citation by Force11 Working group: Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science, 2015.
17. Experimental Metadata Model - preliminary work to model metadata about the experiments that produce datasets; collaboration between Elsevier and Oregon Health and Science University.
18. WHO Dataset from International Clinical Trials Registry Program.
19. VIVO-ISF linking people to scholarly products; it is being aligned and integrated with SCIENCV NIH biosketch system.
20. ISO/IEC JTC1 SC32 WG2: Working Group that develops international standards for metadata and related technologies.
21. CERIF and EuroCRIS models.
22. Provenance, Annotation and Versioning (PAV) ontology.

2.1.2. Metamodels

This is a non-comprehensive list of the metamodels analysed, which might have more specific scopes and different use cases than the intended capability of the DataMed prototype.

1. ISO/IEC 11179: Metadata Registries - Part 3: Registry metamodel and basic attributes.
2. ANSI X3.285: Metamodel for the Management of Shareable Data: conceptual model for the specification of a data registry
3. DataFairport Profiles.
4. Research Object Ontology (based on OAI-ORE for aggregation, W3C Web Annotation Data Model for annotations and W3C PROV for provenance).
5. Minimum Information Model ontology - a metamodel for describing minimum information model.

2.2. Top-down Use Cases

Use cases have been guiding elements throughout the process, in order to define the appropriate boundaries and level of granularity: which queries will be answered in full by the DataMed prototype, which only partially, and which are out of scope.

2.2.1. Extracting Competency Questions

The use cases have been: (i) collected at the bioCADDIE Use Cases Workshop, (ii) extracted from the bioCADDIE White Paper, (iii) submitted by the community, and (iv) provided by the NIH to the bioCADDIE Executive Committee. This section describes the methods used to analyse the use cases and derive information on the type of metadata elements needed to support them. From the use cases, a set ‘competency questions’ have been derived; these are defined as the questions which we want the NIH BD2K DataMed prototype to be able to provide support for.

Subsequently the questions have been abstracted, key concepts highlighted and color-coded and binned in entities, attributes and values categories, to be easily matched with the result of the ‘bottom-up approach’. The questions below are grouped according to their source, using an internal code for tracking propose only.

Internal	Competency question

bioCADDIE code	
BGUC1-1	Search for disease x data of all types across all databases (Note: these first three use cases are linked; also there is a Common Data Element for the disease x [HD])
BGUC1-2	Search for data type x related to disease x and disease y to compare behavioral studies (HD and ADHD)
BGUC1-3	Search for data on diseases c, d, e, and f that mention disease x or the disease x gene
BGUC2	Search for organism x in biological process y (apoptosis) at scale z with an estimate of the reliability of the annotations
BGUC3-1	Search for new drug x to predict and track biological process y (cardiotoxicity)
BGUC3-2	Search for data type x ('omics correlates) of biological process for drugs related to drug x
BGUC3-3	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
BGUC3-4	Track responses to treatment to ensure detection of biological process x
BGUC3-5	Find patient data *"like these"* with similar treatments, responses to treatment, genetics
BGUC4	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.
BGUC5	Search for patient data with identifiers linked to data type x (genome) and type z (fMRI) to find variants causal for disease x (autism)
BGUC5-1	Search for patient data with permission a, size b, demographic characteristic c, biosamples available, and data type d (e.g., imaging) available
BGUC5-2	Find Publications a-z related to dataset x
BGUC5-3	Search for studies a-z that tested drug x with agent y and agent role z
BGUC5-4	Search for data on adverse outcome x (obesity as measured by BMI) and disease y (e.g., diabetes) using standard z with license a and quality indicator b and provenance c
BGUC5-5	Search for data that was subsetted based on vaccination *history*
BGUC5-6	Search for data by NIH researchers with > 100 publications on disease x that were peer reviewed
BGUC5-7	Search for data that were curated according to standard x by researcher y or project z
BGUC5-8	Search for data that can be redistributed for free under license x
BGUC5-9	Search for substance x in groundwater to correlate with outcomes in patients with disease z family *history*
BGUC5-10	Search for patients with phenotype x and disorder y (e.g., > 4 drinks a day)

BGUC5-11	Search for **patients** with **exposure to substance** x correlated with **biological process** (mutation) in **genes** a-z
PB1	Search for **data type** x (**gene expression**) analysis on **mouse red blood cells** and narrow search results by **access statistics**
PB2	After search determine which data in result set are **most relevant**

SPUC1	Search for **birth cohort** x (**adolescents**) with combination of **imaging data types** a-z to identify **phenotypes** a-z predictive of **disorders** x and y (**alcohol and drug use**)
SPUC2	Search for **data type** x (imaging data), ***across the lifespan***, with **deep phenotyping** and **data type** y (GWS data)
SPUC3 **PRE3**	Search for **birth cohort data** that are **harmonized** on **variable** x (**educational attainment**) to understand ***historical*** impact on **biological process** y (**adult mortality**)
SPUC4	Query broader and updated **phenotypic categories** for generalized enrichment analysis on **data type** ('omics)
SPUC5	Create **virtual networking environment**, linking **data types** x and y and **literature** to understand **biological process** (molecular biology of carcinogenic **pathway**), which is **accessible** to **medical professionals and patients**.
SPUC6	Search for constraints of **genotypes** a-z and **phenotypes** a-z
SPUC7-1	Search for **EHR data** to monitor **side effects** of **drug** x with **condition/context** y, **data quality** z, prevalence of **medication use**, etc.
SPUC7-2	Link **EHR data** with **knowledge bases** a-z (e.g., SemMedDB, DrugBank, etc.)
SPUC7-3	Search for **clinical patient data** over the course of **disease** x to study **disease progression**, **treatment change and discontinuation, outcomes, condition** (hospital setting)
SPUC8	Search for **longitudinal survey data** on **disorder** x (e.g., tobacco use) with **data type** y (biomarkers)
SPUC9	Search for **patterns** indicative of **drug response** in the **genome** and **transcriptome** with documented **experimental conditions**
SPUC10	Search for **patients** with **disorder** x (e.g., autism) and with specific **data type** (genomic, microbiome and sensor data) profiles; export to **big data compute platform**.
SPUC11	Search for **code snippets** in **statistical software package** x to **extract** or **combine** specific **variables**
SPUC12	Limit searches to **datasets** with different **access requirements** (e.g., IRB, DUA, public)
SPUC13	Search for candidate **genes** a-z associated with **biological process** x (aging) and validate them
SPUC14	Search for **drug-drug interactions** through **automated extraction of structured**

	metadata in an RDF nanopublication and cite associated paper x
SPUC15	Search for **patient data** from **multiple clinical trials** (in academia and industry, with unique IDs for each **clinical trial** and **datasets** within them) to combine them
SPUC16	Search for **datasets a-z** relevant to **causal analysis** in **domains a-z** for use with causal discovery algorithms
SPUC17	Search for ***life histories*** with **data type** x (clinical) on outcomes of **biological process x** (pregnancy) in women with **disorder x** (Factor 11 deficiency)
SPUC18	Search for **pathway** x that regulates at least two of the **genes** in response to **cell stress** x (e.g., UPR)
SPUC19	Search for **clinical trials data** with **policies** x and y (to study transparency)

WPUC1	Search for **patients** with **disease** x (Alzheimers) that have **data types** x, y, and z **available** (e.g., RNA-seq, behavioral, **imaging**)
WPUC2	Search for **data types** x and y related to the same **biological process** z
WPUC3	Search for **data types** x (genome data) with **biological process** (mutations) y and z in **species/organism** a for **phenotype** b
WPUC4	Search for **data elements** and **instruments** that measure **biological process** x (stress); use facets to find different types of **stressors**
WPUC5-p7	Search for **dataset** x **referenced** in **paper** y and determine if **dataset** x is the ***latest*** **version**
WPUC6-p7	What **genes** are differentially expressed in the **ureteric bud** vs. **the mesonephric duct**? (can be derived from a computation -- will such services be connected?)
WPUC7-p7	Search for **datasets** published as a result of **grant** x (how many?)
WPUC8-p7	Search for **datasets** produced from **funder** x (NIH) (how many?)
WPUC9-p7	Search for number of times **gene expression** x (GSE3114) has been analyzed; is it **available** in **format** y?
WPUC10-p7	Which **datasets** funded by **funder** x generated the **most publications?**

UC2	Search for **data** from **author** x, from **database** y, linked to **publication** z
UC15	Search **MIAME standard** compliant **data**, from **database x**
UC1	Search for **data type** x (**gene expression**) in **human cell line** x, funded by **funder** x

2.2.2. Identify Entities Attributes and Values

The concepts highlighted in the use cases above have been binned in entities, attributes and values categories.

material entity

Organization/
NIH[BGUC5-6,WPUC8-p7]
Biomaterial/
human cell line [UC1]
organism x [BGUC2,WPUC3]
mouse [PB1]
human/Homo sapiens [UC1]
population/cohort [SPUC1,SPUC3]
family [BGUC5-9]
BGUC5 [BGUC5-1]
groundwater [BGUC5-9]
red blood cells [PB1]
ureteric bud [WPUC6-p7]
mesonephric duct [WPUC6-p7]
Molecular entity/
gene [BGUC1-3,BGU5-11,SPUC13,SPUC18,WPUC6-p7]
protein {placeholder}
nucleic acid{placeholder}
metabolite {placeholder}
chemical entity
drug/medication [BGUC3-1,BGUC3-2,BGUC3-3,BGUC5-3,SPUC1,SPUC7-1,SPUC14]
Material entity
Instrument [WPUC4]
Process
Biological process/ [WPUC2,WPUC3,WPUC4, SPUC5, SPUC13,SPUC17,BGUC5-11, BGUC2,BGUC3-1,BGUC3-2,BGUC3-4,BGUC4]
gene expression [PB1,UC1,WPUC9-p7]
disease progression [SPUC7-3]
cell stress [SPUC18]
mutation [WPUC3]
Planned Process
peer-review [BGUC5-6]
curation [BGUC5-7]
publishing [WPUC7-p7]
distributing [BGUC5-8]
imaging [SPUC1,SPUC2,WPUC1]
referencing/citing [WPUC5-p7, SPUC14]
Study
longitudinal survey [SPUC8]
clinical trials [SPUC15,SPUC19]
intervention/experimental condition/stressor/treatment[BGUC3-4,BGUC3-5,WPUC4, SPUC9,SPUC7-1(*)]
vaccination [BGUC5-5]
analysis/data transformation
generalized enrichment analysis [SPUC4]
causal analysis [SPUC16]
harmonization [SPUC3]
differential analysis [WPUC6-p7]
correlation analysis [BGUC5-9,BGUC5-11]
Unplanned Process
Adverse event / Side effect [SPUC7-1]
Property
role/[BGUC5-3]
researcher [BGUC5-7]
author [UC2]
funder [WPUC8-p7,WPUC10-p7, UC1]
medical professionals[SPUC5]

patient [WPUC1,SPUC15,SPUC10,SPUC7-3,SPUC5, BGUC4, BGUC5-9,BGUC5-10,BGUC5-11, BGUC3-3,BGUC3-5,BGUC5,BGUC5-1]
developmental stage
adolescent [SPCU1]
adult [SPUC3]
Phenotype/ [BGUC5-10,WPUC3, SPUC6,SPUC1]
    demographic characteristic [BGUC4,BGUC5-1]
    phenotypic categories [SPUC4]
Disease/ [BGUC1-1,BGUC1-2,BGUC1-3,BGUC5, BGUC5-4,BGUC5-6,BGUC-5-9,SPUC7-3,WPUC1]
    disorder [SPUC1,SPUC-8,SPUC10,SPUC17, BGUC5-10]
    obesity [BGUC5-4,BGUC4]
    autism [BGUC5]
mortality [SPUC3]
availability [BGUC5-1, SPUC5]
quality [SPUC7-1, BGUC5-4]
reliability [BGUC2]
relevance [PB2]
similarity [BGUC3-2,BGUC3-3,BGUC3-5]
compliance [UC15]
provenance [BGUC5-4]
prevalence [SPUC7-1]
Information content entity
Bioinformatic Resource
    knowledge base [SPUC7-2]
statistical software package [SPUC11]
big data compute platform [SPUC10]
pathway [SPUC5,SPUC18]?
identifier [BGUC5, SPUC14]
Publication [UC2-9]
    annotation [BGUC2]
    literature [UC27]
    paper/publication [BGUC5-2,BGUC5-6,SPUC14, UC2,WPUC10-p7]
Specification/Collection of Rules
format [WPUC9-p7]
standard [U15,BGUC5-4,BGUC5-7]
license [BGUC5-4,BGUC5-8]
policy [SPUC19]
permission [BGUC5-1]
version [WPUC5-p7]
Data/Measurement
    gene expression data [uc1]
    imaging data [SPUC1,SPUC2]
deep phenotyping and GWS data [SPUC2]
birth cohort data [SPUC3/PRE3]
genome data [BGUC5,WPUC3]
fMRI data [BGUC5]
omics data [uc26]
eHR data [uc28]
variable [SPUC3,SPUC11]
    educational attainment [SPUC3]
scale [BGUC2]
size [BGUC5-1]
Temporal interval
History [BGUC5-9, BGUC5-5, BGUC3-3,SPUC3,SPUC17,WPUC5-p7]
lifespan [SPUC2]

2.3. Bottom-up Mapping

The generic metadata schemas and some life science-specific one that have been mapped to identify common metadata elements are in a separate ***"Appendix IV - NIH BD2K bioCADDIE WG3 Metadata Mapping File v1.1"*** is available from the bioCADDIE Github repository.

2.3.1. Input Material

Generic metadata schemas and some life science-specific ones have been mapped to identify common metadata elements. When available, formal representations such XML schema document (XSD) and semantic model (RDF/OWL representations) have been used as input material in the mapping process. The mapping in Appendix III covers the schemas listed below, encompassing both generics and science-specific schemas.

2.3.1.1. Generic Metadata Schemas and Models

1. Datacite Metadata Schema
2. Schema.org; Dataset class, a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet; used by search engines such as Google, Bing and Yahoo.
    a. The Dataset class in Schema.org, used in this WG mapping, is based upon the W3C Data Catalog Vocabulary (DCAT) and it benefits from collaboration around the DCAT, ADMS and VoID vocabularies; details and mappings.
3. Dataset Descriptions: W3C HCLS Community Profile. 2015
    a. Dataset descriptors identification file
4. NLM preliminary work on metadata core set
5. Registry Interchange Format Collections and Services (RIF-CS), used in the JISC Research Data Registry and Discovery Service
    a. Documentation
    b. Implementation of a profile of ISO 2146.
6. Project Open Data Metadata Schema v1.1

2.3.1.2. Life Science Metadata Schemas

1. NCBI BioProject / NCBI BioSample
2. EMBL-EBI Pride.xsd
3. EMBL-EBI/NCBI Short Read Archive xsd
4. Nature's *Scientific Data* ISA specification and ISA file (study metadata) as ingested in the article XML
5. EMBL-EBI MetaboLights ISA configuration
6. NCBI Gene Expression Omnibus MiniML.xsd
7. CDISC BRIDG Model 3
8. CDISC SDM.xsd, which imports CDISC ODM.xsd
9. GA4GH metadata model

In addition, existing mapping and comparisons has also been reviewed and considered:

1. An initial comparison by bioCADDIE Development Team's members.
2. linkedISA experimental metadata mapped to OBO Foundry OBI and the Semantic science Integrated Ontology (SIO)

2.3.2. BioSharing Collection of Schemas/Models

The metadata schemas and models used in the mapping have been described in the BioSharing Collection for bioCADDIE, which will be enriched progressively; the information includes:

- creators and maintainers;
- documentation, including URL where this is located;

and when available

- version;
- source of metadata elements (e.g. XSD), including the URL where the model or schema has been sourced.