NIH BD2K bioCADDIE WG3 - DataMed Data Discovery Index

**DATS Metadata Specification v1.1**

**Status:**

# **Working document version 1.1**

*Stable version v 1.1 release schedule: May 2016.*

**Working Group Goals:**

The [NIH BD2K bioCADDIE Metadata Working Group (WG) 3](https://biocaddie.org/group/working-group/working-group-3-metadata-specifications) is tasked to define a set of metadata specifications that support intended capability of the [NIH BD2K Data Discovery Index prototype](https://datascience.nih.gov/bd2k/funded-programs/resource-indexing), named **DataMed,** as outlined in the [bioCADDIE White Paper](https://biocaddie.org/publications/biocaddie-white-paper). The specifications encompass: (i) a **model**, named **DATS** (DatA Tag Suite), to describe the metadata and the structure for datasets along the line of the [JATS](http://jats.nlm.nih.gov/index.html) used by PubMed for literature; and (ii) its **serializations** in JSON and/or other formats. In a subsequent phased, this group will explore mapping the DATS model to schema.org to define an extension as part of the [bioschemas.org](http://bioschemas.org/) initiative. The Metadata WG3 is a joint activity with [bioCADDIE WG7](https://biocaddie.org/group/working-group/working-group-7-accessibility-metadata-datasets) and [CEDAR](http://metadatacenter.org/), a BD2K centre focused on metadata, and closely connected to to other NIH initiatives and [ELIXIR](https://www.elixir-europe.org/) activities in Europe. This joint Metadata WG3 has produced a ‘core metadata’ designed to be future-proofed for progressive extensions to accommodate ‘domain-specific metadata’ for more specialized data types as needed.

**Scope of the Document:**

This document describes: (i) the process leading to v1.0, the material reviewed (section 1 and Appendix I) and the use cases (section 2) used to identify an initial set of metadata elements and create a JSON schemata (section 3 and Appendix II); the changes addressed to deliver the current DATS specification v1.1.

**Associated Material:**

Appendix I (Metadata Mapping File v1.1) and Appendix II (DATS Metadata Elements File v1.1), along with the JSON schemata are available from the [bioCADDIE Github repository](https://github.com/biocaddie/WG3-MetadataSpecifications) (under the Version 1.1 section). The [bioCADDIE Metadata WG3](https://biocaddie.org/workgroup-3-group-links) webpage has links to presentations and notes from the WG3 activities. The schemas and models in Appendix I are also listed in the [BioSharing Collection for bioCADDIE](https://www.biosharing.org/collection/BioCaddie).

**Intended Audience:**

This document is aimed at: (i) the [bioCADDIE Core Development Team](https://biocaddie.org/core-development-team) that will implement and test this model, (ii) prospective data sources that wish to be indexed in DataMed, and (iii) developers of data harvesting and other metadata tools.

**Contact:**

Comments can also be added to this Google document, or sent to Susanna-Assunta Sansone (WG3 chair) and [WG3 members](https://biocaddie.org/node/392/members) via [biocaddie[at]ucsd.edu](mailto:biocaddie@ucsd.edu)(subject line: WG3 Spec V1.1)

**Table of contents:**

[1. Standard Operating Procedure](#h.f2unvbebrq7f)

[1.1. Combined Approaches](#h.yquzapx4kwvd)

[1.1.1. Data Discovery Initiatives and Metadata Initiatives](#h.8mzvjltr2f6f)

[1.1.2. Metamodels](#h.u5yqo01cj5ce)

[1.2. Top-down Use Cases](#h.rr21ap8c7nwo)

[1.3. Bottom-up Mapping](#h.wdhs1ilc10i0)

[1.3.1. Input Material](#h.g3tnetuf743u)

[1.3.1.1. Generic Metadata Schemas and Models](#h.iw8jj353tf8g)

[1.3.1.2. Life Science Metadata Schemas](#h.y1p2gd49eq3f)

[1.3.2. BioSharing Collection of Schemas and Models](#h.de50i5ry2mzb)

[2. Use Cases and Derived Metadata](#h.ms3zn25kdpn8)

[2.1. Methodology](#h.tjwgzmx1g6j6)

[2.1.1. Competency Questions](#h.sg1uk9w8lax2)

[2.1.2. Entities Attributes and Values](#h.pyn0a055jkqe)

[3. The DATS Metadata Elements](#h.1iu4um19v9z8)

[3.1. Overview and General Considerations](#h.6otc4ztw6ma9)

[3.2. Evaluation and Review of v1.0 and Creation of v1.1](#h.uia27s2awq12)

[3.3. JSON Schema v1.1](#h.pkus9ucyfcf)

[3.4. Detailed Description of v1.1](#h.h6penuyw719p)

[3.4.1. Cardinality and Requirement Level](#h.7gg35ke4gczq)

[3.4.2. Links to Use Cases and Schemas/Models](#h.bkyez7wixoyc)

[4. Appendix I](#h.y64tep1umthr)

[5. Appendix II](#h.v8wcbsqppc5k)

# 

# 

# 1. Standard Operating Procedure

***Authors: Alejandra Gonzalez-Beltran******[[1]](#footnote-0), Philippe Rocca-Serra1, Susanna-Assunta Sansone1 and WG3 members.***

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

## 1.1. 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 (section 1.1.1). Several metamodels for representing metadata also exist and have been reviewed (section 1.1.2). 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 1.2); and
* a mapping of existing metadata schemas (bottom-up approach; section 1.3 and Appendix I).

### 1.1.1. 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 NIH BD2K DataMed prototype.

1. UK [JISC Research Data Registry and Discovery Service](http://www.dcc.ac.uk/projects/research-data-registry-pilot): relies on Registry Interchange Format Collections and Services ([RIF-CS](http://guides.ands.org.au/rda-cpg/rifcs)); related documentations: [Github repository](https://github.com/DigitalCurationCentre), [WP3: Metadata Development and Standardisation](http://www.dcc.ac.uk/sites/default/files/documents/registry/uk-rdr-wp3-report-v01.pdf), [Report: metadata mapping schemes / recommendations (version 9, 2014-05-09).](http://www.dcc.ac.uk/projects/research-data-registry-pilot#sthash.6wpCmBTl.dpuf)
2. [Datacite Metadata Search](http://search.datacite.org/ui) to search datasets registered with Datacite.
3. European [EUDAT B2FIND](http://b2find.eudat.eu/): relies in [CKAN](http://docs.ckan.org/en/ckan-1.8/domain-model-dataset.html) ([CKAN Dataset Model](http://docs.ckan.org/en/ckan-1.8/domain-model-dataset.html)) and harvest data using the Open Archives Initiative Protocol for Metadata Harvesting ([OAI-PMH](https://www.openarchives.org/pmh/)). Other references: [documentation](http://eudat.eu/User%20Documentation%20-%20B2FIND.html) and [mapping files](https://github.com/EUDAT-B2FIND/md-mapping).
4. Research Data Alliance (RDA) [Research Data Switchboard](http://www.rd-switchboard.org/) relies on OAI-PMH protocol ([Github repository](https://github.com/rd-alliance-ddri/research-data-switchboard)).
5. [National Data Service](http://www.nationaldataservice.org/).
6. National Institute of Health’s Neuroscience Blueprint funded [Neuroscience Information Framework (NIF).](http://neuinfo.org)
7. The National Institute of Diabetes and Digestive and Kidney Diseases’ [NIDDK Information Network (dkNET).](http://dknet.org)
8. [Data Documentation Initiative](http://www.ddialliance.org/) Draft Specification of [DDI-RDF Discovery Vocabulary](http://rdf-vocabulary.ddialliance.org/discovery.html)(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](http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CCIQFjAAahUKEwjtrP2ny5bGAhUondsKHZR8AKA&url=http%3A%2F%2Fwww.w3.org%2FTR%2Fvocab-data-cube%2F&ei=HVWBVa2qFqi67gaU-YGACg&usg=AFQjCNFVGQ1Hqncw14ceQYNtajfQkZe_2Q&sig2=GeYabhlE1EJKwIIrSBSg3w&bvm=bv.96041959,d.ZGU), [DCAT](http://www.w3.org/TR/vocab-dcat/) and [XKOS](https://raw.githubusercontent.com/linked-statistics/xkos/master/xkos.ttl)).
9. [Global Alliance for Genomic Health](http://genomicsandhealth.org/): specifically [Data Working Group](http://ga4gh.org/), [Genotype-2-phenotype task team](http://genomicsandhealth.org/working-groups/our-work/genotype2phenotype-association), NIH Office of Director funded [Monarch Initiative](http://monarchinitiative.org/) - these are all efforts to develop standardized schemas for genotype-phenotype data integration and data sharing.
10. [eTRIKS standards starter pack](http://www.etriks.org/project/etriks-standards-starter-pack/), under a pre-competitive private, public Innovative Medicine Initiative - aiming to bridge clinical/CDISC, [ISA](http://isa-tools.org) and other community-based standards.
11. [RDA Working Group on Data Description Registry Interoperability.](https://rd-alliance.org/groups/data-description-registry-interoperability.html)
12. [EBI RDF Platform.](https://www.ebi.ac.uk/rdf/platform)
13. Content Standard for Digital Geospatial Metadata [Part 1: Biological Data Profile](http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/biometadata/biodatap.pdf), 1999.
14. [Open PHACTS](http://openphacts.org/) Discovery Platform - integrating pharmacological data resources according to [Dataset descriptions for the Open Pharmacological Space](http://www.openphacts.org/specs/2013/WD-datadesc-20130912/), based on W3C [VoID](http://www.w3.org/TR/void/) for describing Linked Datasets, [HCLS Dataset descriptions](http://www.w3.org/TR/hcls-dataset/) by the W3C Semantic Web in Health Care and Life Sciences Interest Group.
15. [Just Enough Results Model (JERM)](http://seek4science.org/jerm) from the [SEEK for Science](http://seek4science.org/) project.
16. Metadata for data citation by Force11 Working group: [Achieving human and machine accessibility of cited data in scholarly publications](https://dx.doi.org/10.7717/peerj-cs.1). PeerJ Computer Science, 2015.
17. [Experimental Metadata Model](https://github.com/OHSU-Ontology-Development-Group/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](http://www.who.int/ictrp/network/trds/en/) from International Clinical Trials Registry Program.
19. [VIVO-ISF](https://github.com/vivo-isf) linking people to scholarly products; it is being aligned and integrated with [SCIENCV NIH](http://www.ncbi.nlm.nih.gov/sciencv/) biosketch system.
20. [ISO/IEC JTC1 SC32 WG2](http://metadata-standards.org/): Working Group that develops international standards for metadata and related technologies.
21. CERIF and [EuroCRIS](http://www.eurocris.org/) models.
22. [Provenance, Annotation and Versioning (PAV) ontology.](http://www.jbiomedsem.com/content/4/1/37)

### 1.1.2. 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 NIH BD2K DataMed prototype.

1. [ISO/IEC 11179: Metadata Registries](http://metadata-stds.org/11179/) - Part 3: Registry metamodel and basic attributes.
2. [ANSI X3.285: Metamodel for the Management of Shareable Data](http://metadata-standards.org/Document-library/Draft-standards/X3-285-Mgmt-of-Sharable-Data/X3-285.PDF): conceptual model for the specification of a data registry
3. [DataFairport Profiles.](https://github.com/FAIRDataInitiative/DataFairPort)
4. [Research Object Ontology](https://w3id.org/ro#ro) (based on [OAI-ORE for aggregation](https://www.openarchives.org/ore/), [W3C Web Annotation Data Model for annotations](http://www.w3.org/TR/annotation-model/) and [W3C PROV](http://www.w3.org/TR/prov-overview/) for provenance).
5. [Minimum Information Model ontology](https://github.com/wf4ever/ro-manager/blob/master/Minim/Minim-description.md) - a metamodel for describing minimum information model.

## 

## 1.2. 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 NIH BD2K DataMed prototype, which only partially, and which are out of scope. From a selection of competency questions derived from different sources key metadata elements have been highlighted and color-coded to be easily matched with the metadata resulting from the bottom-up mapping approach, described below. This top-down approach is detailed in section 2.

## 1.3. Bottom-up Mapping

### 1.3.1. 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 is available as Appendix I - NIH BD2K bioCADDIE WG3 Metadata Mapping v1.1 (section 4). The mapping covers the schemas listed below, encompassing both generics and science-specific schemas.

#### **1.3.1.1. Generic Metadata Schemas and Models**

1. [Datacite Metadata Schema](https://www.biosharing.org/bsg-000588)
2. [Schema.org](https://www.biosharing.org/bsg-000593); [Dataset class](https://schema.org/Dataset), 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.
   1. The Dataset class in Schema.org, used in this WG mapping, is based upon the [W3C Data Catalog Vocabulary (DCAT)](http://www.w3.org/TR/vocab-dcat/) and it benefits from collaboration around the DCAT, ADMS and VoID vocabularies; [details and mappings](http://www.w3.org/wiki/WebSchemas/Datasets).
3. [Dataset Descriptions: W3C HCLS Community Profile. 2015](https://www.biosharing.org/bsg-000579)
   1. [Dataset descriptors identification file](https://docs.google.com/spreadsheets/d/1bhbw1HAp5I_c9JvAxyURKW0uEGJ8jV5hlzf07ggWDxc/edit#gid=1)
4. [NLM preliminary work on metadata core set](https://www.biosharing.org/bsg-000600)
5. [Registry Interchange Format Collections and Services (RIF-CS](https://www.biosharing.org/bsg-000591)), used in the JISC Research Data Registry and Discovery Service
   1. [Documentation](http://guides.ands.org.au/rda-cpg/rifcs)
   2. Implementation of a profile of [ISO 2146](http://www.iso.org/iso/catalogue_detail.htm?csnumber=44936).
6. [Project Open Data Metadata Schema v1.1](https://project-open-data.cio.gov/v1.1/schema/)

#### **1.3.1.2. Life Science Metadata Schemas**

1. [NCBI BioProject](https://www.biosharing.org/bsg-000598) / NCBI BioSample
2. [EMBL-EBI Pride.xsd](https://www.biosharing.org/bsg-000561)
3. [EMBL-EBI/NCBI Short Read Archive xsd](https://www.biosharing.org/bsg-000084)
4. Nature’s *Scientific Data*[ISA specification](http://www.nature.com/uploads/ckeditor/attachments/1377/NPG_DD_spec_v1b_July2014.pdf)and [ISA file (study metadata) as ingested in the article XML](http://www.ncbi.nlm.nih.gov/books/NBK279831/)
5. EMBL-EBI MetaboLights [ISA configuration](https://github.com/ISA-tools/Configuration-Files/tree/master/isaconfig-default_v2015-07-02)
6. [NCBI Gene Expression Omnibus MiniML.xsd](https://www.biosharing.org/bsg-000076)
7. [CDISC BRIDG Model 3](https://www.biosharing.org/bsg-000597)
8. [CDISC SDM.xsd](https://www.biosharing.org/bsg-000566), which imports CDISC [ODM.xsd](https://www.biosharing.org/bsg-000566)
9. [GA4GH metadata model](https://www.biosharing.org/bsg-000599)

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

1. An [initial comparison by bioCADDIE Development Team’s members](https://drive.google.com/open?id=0B1V2WmAsn-OkflVIbm5VU0FwaFNhZ0VsOVBGUU5PSVp5VVFvdzVRYnpjcHVtN3VqNDBhak0&authuser=0).
2. [linkedISA experimental metadata](http://www.biomedcentral.com/1471-2105/15/S14/S4) mapped to OBO Foundry OBI and the [Semantic science Integrated Ontology (SIO)](http://sio.semanticscience.org/)

### 1.3.2. BioSharing Collection of Schemas and Models

The metadata schemas and models used in the mapping have been described in the [BioSharing Collection for bioCADDIE](https://www.biosharing.org/collection/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.

# 

# 2. Use Cases and Derived Metadata

***Authors: Philippe Rocca-Serra***[***1***](#id.va8c02egru63)***, Mary Vardigan[[2]](#footnote-1) and WG3 members.***

As outlined in section 1, the analysis of the use cases is referred to as top-down approach that - combined to the bottom-up approach (section 1.3 and Appendix I) - has been used to identify the initiate set of metadata elements. The use cases have been: (i) collected at the [bioCADDIE Use Cases Workshop](https://biocaddie.org/biocaddie-use-case-workshop-invitation), (ii) extracted from the [bioCADDIE White Paper](https://biocaddie.org/publications/biocaddie-white-paper), (iii) submitted by the community, and (iv) provided by the NIH to the [bioCADDIE Executive Committee](https://biocaddie.org/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.

## 2.1. Methodology

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’.

### 2.1.1. Competency Questions

The questions below are grouped according to their source, using an internal code for tracking propose only.

|  |  |
| --- | --- |
| **Internal**  **bioCADDIE code** | **Competency question** |
| **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 withoutcomes 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** xcorrelated 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) inwomen with **disorder x** (Factor 11 deficiency) |
| **SPUC18** | Search for **pathway** x that regulates at least two of the **gene**s 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 **gene**s 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 |

# 

### 

### 2.1.2. Entities Attributes and Values

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

|  |
| --- |
| [**material entity**](http://bioportal.bioontology.org/ontologies/OGG?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FBFO_0000040)**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]** |

# 

# 3. The DATS Metadata Elements

***Authors: Alejandra Gonzalez-Beltran***[***1***](#id.va8c02egru63)***, Philippe Rocca-Serra***[***1***](#id.va8c02egru63)***, George Alter***[***2***](#id.l9gzgrxhpv1)***, Susanna-Assunta Sansone***[***1***](#id.va8c02egru63)***, WG3 and WG7 members.***

This section describes the initial set of DATS metadata elements derived from the bottom-up approach - as they were present in more than one of the schemas analysed (section 1.3 and Appendix I) - and the use cases-driven top-down approach (section 2). An high level overview of the key metadata elements and and their relations is provided (section 3.1), along with a detailed description of each entity (section 3.2 and Appendix II).

## 3.1. Overview and General Considerations

The DATS model is also made available as machine readable JSON schemata; an overview of some of the metadata elements, their types and relations is shown in the Figure 1. All elements, along with their definition, relations and further details are described in section 3.3. This overview also illustrates that the model is designed around the *Dataset* object. This entity is also linked to other digital research objects part of the [NIH Commons](https://datascience.nih.gov/commons), such as *Software* and *Data Standard*, which are the focus on other discovery indexes and therefore 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, 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 competency questions can be answered in full and that these may not be representative of all kind of data sources the NIH BD2K DataMed prototype should retrieve information from.

## 3.2. Evaluation and Review of v1.0 and Creation of v1.1

In August 2015, the specification [v1.0 was released (DOI:10.5281/zenodo.28019)](http://dx.doi.org/10.5281/zenodo.28019). These initial metadata elements were formally represented - adopting the [FAIR principles](http://www.nature.com/articles/sdata201618) - and tested by the [bioCADDIE Core Development Team](https://biocaddie.org/core-development-team) with a variety of sources. Following this evaluation and the review phase by the larger community, several changes and additions have been made to v1.0 to deliver the current DATS specification v1.1. The edits include, but are not limited to:

* revision of the requirement levels
* better use of the *identifier* entity and addition of more properties
* revision of the *role* property
* creation of *data* entity
* generalization of *type* property
* expansion of *version* property
* distinctions between different types of data sources
* generic element to support extra metadata
* addition of [accessibility metadata elements produced by bioCADDIE WG7](https://biocaddie.org/group/working-group/working-group-7-accessibility-metadata-datasets).

For more details and pending edits, see the ***Working Note*** section at the end of this document.

**Figure 1.** A schematic overview of some of the DATS metadata elements in v1.1, their types and relations.

### 

## 3.3. JSON Schema v1.1

The DATS metadata elements are also available as machine readable JSON schemata from [bioCADDIE Github repository](https://github.com/biocaddie/WG3-MetadataSpecifications/tree/master/json-schemas).

## 3.4. Detailed Description of v1.1

A full description of the DATS metadata elements, grouped by types and color coded is available as a separate Appendix II - NIH BD2K BioCADDIE WG3 DATS Metadata Elements File v1.1 (section 5); the [Google JSON style guide](http://google-styleguide.googlecode.com/svn/trunk/jsoncstyleguide.xml) 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); others are detailed below.

### 3.4.1. Cardinality and Requirement Level

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 key words "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](http://json-schema.org/latest/json-schema-core.html#RFC2119) [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 iteratively reviewed and used to evaluate if a data/database source ‘complies’ to the DATS metadata, and therefore if it has the potential to fulfil the relevant competency questions.

### 3.4.2. Links to Use Cases and Schemas/Models

The DATS metadata elements are associated to relevant use cases and competency questions (section 2.1.1). In Appendix II, for those entities and/or properties were no specific competency questions are indicated, links to the relevant schema(s)/model(s) are also provided, to justify their relevance and provenance.

# 4. Appendix I

***Authors: Alejandra Gonzalez-Beltran***[***1***](#id.va8c02egru63)***, Philippe Rocca-Serra***[***1***](#id.va8c02egru63) ***and WG3 members.***

***NIH BD2K bioCADDIE WG3 Metadata Mapping File v1.1***, is available from the [bioCADDIE Github repository](https://github.com/biocaddie/WG3-MetadataSpecifications/blob/master/AppendixI-WG3MetadataMappingFilev1-NIH-BD2K-bioCADDIE-DataDiscoveryIndex.xlsx). The file describes the generic metadata schemas and some life science-specific one that have been mapped to identify common metadata elements.

# 5. Appendix II

***Authors: Alejandra Gonzalez-Beltran***[***1***](#id.va8c02egru63)***, Philippe Rocca-Serra***[***1***](#id.va8c02egru63)***, George Alter***[***2***](#id.l9gzgrxhpv1)***, Susanna-Assunta Sansone***[***1***](#id.va8c02egru63)***, WG3 and WG7 members.***

***NIH BD2K bioCADDIE WG3 DATS Metadata Elements File v1.1*** is available as a [Google spreadsheet here](https://docs.google.com/spreadsheets/d/1aHj_Qvlr7Sf4DlU4uc37PQEOPha8jTxMEdyL_Q7geLQ/edit#gid=0), and also linked availbale from the [bioCADDIE Github repository](https://github.com/biocaddie/WG3-MetadataSpecifications) (under the Version 1.1 section). The file describes the metadata elements, grouped by types, providing details on their: definition, attributes, requirements, cardinality and requirement level. Notes in the spreadsheet also indicate areas that are still being addressed, also detailed in the ***Working Notes*** at the end of this document.

**Working Notes**

Comments received and being addressed are listed here (grouped by status):

* **Addressed** 
  + *homepage* is removed (from *DataStandard* and *DataRepository* entities) and will be addressed via the extended identifier entity
  + added Person.*fullName* and Person.*midInitial* properties, supporting all options for ingesting datasets from different sources (that may or not have separate fields for each section of the name)
  + [accessibility metadata](https://docs.google.com/document/d/15QN34r5BMmxlXpB4bVUIO_mgizQl7VcQjCp80tK5JRQ/edit) elements from bioCADDIE WG7 are added
    - Added an entity for Access to the DATS model
    - Access.accessTypes changed to ‘type’ (to make it consistent with the other entities)
    - Refined definitions of different elements
    - The different types presented in the document (access type, authorization type, etc) are provided as examples, as their definition is assumed to be available in a controlled vocabulary.
    - Added cardinality restrictions to the different properties.
    - Dataset.downloadURL moved to the Access entity in the form of ‘accessURL’
    - Dataset.license removed as it is now considered in Access.
    - Added ‘landingPage’ to Access entity
    - The fields whose values come from a CV were changed from string to ‘string or IRI’
    - As the access metadata covers License, we removed Dataset.license element
    - Changed License property from type string to entity License
  + Dataset:relatedDataset is removed and will be address via extended identifier entity
  + Refined definition of *type* in *DataRepository* to be able to distinguish primary for other type of data sources, e.g. aggregators
    - dataRepositoryDataType changed to ‘type’ to indicate the type of the repository, such as ‘primary’ or ‘aggregator’
    - Added property for ‘aggregates’ referring to DataRepositories that this Repositoy might aggregate, in the case that it is not a ‘primary’ resource.
  + *version* property for *Dataset* to be included
  + *keywords* property for *Dataset* to be included
  + *date* becomes a separate entity
  + *type* is generalized
  + Include reference and map to <https://project-open-data.cio.gov/v1.1/schema/>; <https://github.com/project-open-data>
  + Dataset split in Dataset and DatasetDistribution
  + In all the process entities (e.g. Activity, Study, Treatment) etc, startDate and endDate are represented with the Date entity (whose types will be type=”start date” and type=”end date” respectively)
  + *role* property changed its requirement level from SHOULD to MAY, so that the model can cover a variety of roles if needed, but the most common roles are retrieved from associated entities (e.g. a creator for a Dataset is given in the Dataset entity)
  + *identifier* becomes a separate entity, as well as alternateIdentifiers and relatedIdentifiers - this follows discussion with Joan Starr of Datacite
  + Remove specialised ‘derivesFrom’ relationships in Material, only leaving ‘derivesFrom’ Material, which could be used for Organism or AnatomicalPart
  + Regularize attribute to plural forms when cardinality >1 (e.g. license -> licenses)
  + Replace ‘Biological Process’ with ‘Biological Concept’ to allow any of the Gene Ontology Axis to be specified
  + Added Characteristics attribute to ‘Molecular Entity’
  + Requirement levels reviewed
* **In progress**
  + address ways of supporting any extra metadata, often very resources specific
  + define a generic core to form the DATS
  + follow-up on discussion with OMOP community
  + consider DataCite to Dublin Core Metadata Application Profile (<https://groups.google.com/a/datacite.org/forum/#!forum/dc2map>)
  + Check the core development group transformation files
  + Add comment about the possibility of curation status to Dataset as well as creating a CompositeDataset with the criteria for the composition/aggregation
  + Reference field mappings https://project-open-data.cio.gov/v1.1/metadata-resources/#field-mappings

1. University of Oxford, Oxford e-Research Centre, UK; bioCADDIE and CEDAR partner. [↑](#footnote-ref-0)
2. ICPSR, University of Michigan, USA; bioCADDIE partner. [↑](#footnote-ref-1)