NIH BD2K bioCADDIE Data Discovery Index WG3
Metadata Specification v1

Status:

Version 1 - 12th August 2015

Working Group Goals:

The NIH BD2K bioCADDIE Metadata Working Group (WG) 3 is tasked to define a set of metadata specifications that support intended capability of the NIH BD2K Data Discovery Index prototype - as outlined in the bioCADDIE White Paper. This is a joint activity between bioCADDIE and CEDAR, another BD2K centre focused on metadata. This joint Metadata WG3 operates in two phases to produce a 'core metadata' (phase 1), designed to be future-proofed for progressive extensions to accommodate 'domain-specific metadata' for more specialized data types (phase 2), as needed.

Scope of the Document:

This document describes the process, 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); this work will be iteratively reviewed, following a test phase.

Associated Material:

This document, its appendixes and the associated code are available from the bioCADDIE Github repository and the bioCADDIE Metadata WG3 webpage; the latter also serves the original working files, the presentations and notes from the WG3 activities. The schemas and models in Appendix I are also listed in the BioSharing Collection for bioCADDIE.

Intended Audience:

This document is primarily aimed at the bioCADDIE Core Development Team that will implement and test this model; however this is also informative for other parties such as prospective data sources for the Data Discovery Index and developers of data harvesting and other tools.

Contact:

Questions and comments to Susanna-Assunta Sansone (WG3 chair) and WG3 members - adding them directly to the live Google document or sending them to biocaddie[at]ucsd.edu.

Table of Content:

1. Standard Operating Procedure

Authors: Alejandra Gonzalez-Beltran[1], Philippe Rocca-Serra[1], Susanna-Assunta Sansone[1] and WG3 members.

This section outlines the methods and the process used to identify an initial set of metadata elements.

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 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 Data Discovery Index 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).

[1] University of Oxford, Oxford e-Research Centre, UK; bioCADDIE and CEDAR partner.

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.

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 Data Discovery Index 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.

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 Data Discovery Index 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 one 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 (section 4). The mapping for this phase 1 covers the schemas listed below; additional life science-specific schemas will be considered in phase 2, when domain-specific metadata for more specialized data types are tackled.

1.3.1.1. 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.

1.3.1.2. 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)

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, 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.

1.4. Phase 1 Outputs, Evaluation and Next Steps

The approach described in the sections above has delivered an initial set of metadata elements, which have also been ranked and provisionally associated with a requirement level, as per the RFC 2119; the metadata model is available as machine readable JSON schemata. These outputs are described in section 3.

To evaluate this initial set of metadata elements, the work will continue as part of the bioCADDIE Core Development Team: the metadata elements will formally be represented - adopting the FAIR principles - and tested with several data sources. This initial set of metadata elements will be iteratively reviewed, as needed, and domain specific extensions identified as part of the phase 2 activities of the WG3.

2. Use Cases and Derived Metadata

Authors: Philippe Rocca-Serra[1], Mary Vardigan[2] 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, (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

[2] ICPSR, University of Michigan, USA; bioCADDIE partner.

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 Data Discovery Index 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 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

2.1.2. Entities Attributes and Values

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

material entity
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. Initial Set of Metadata Elements

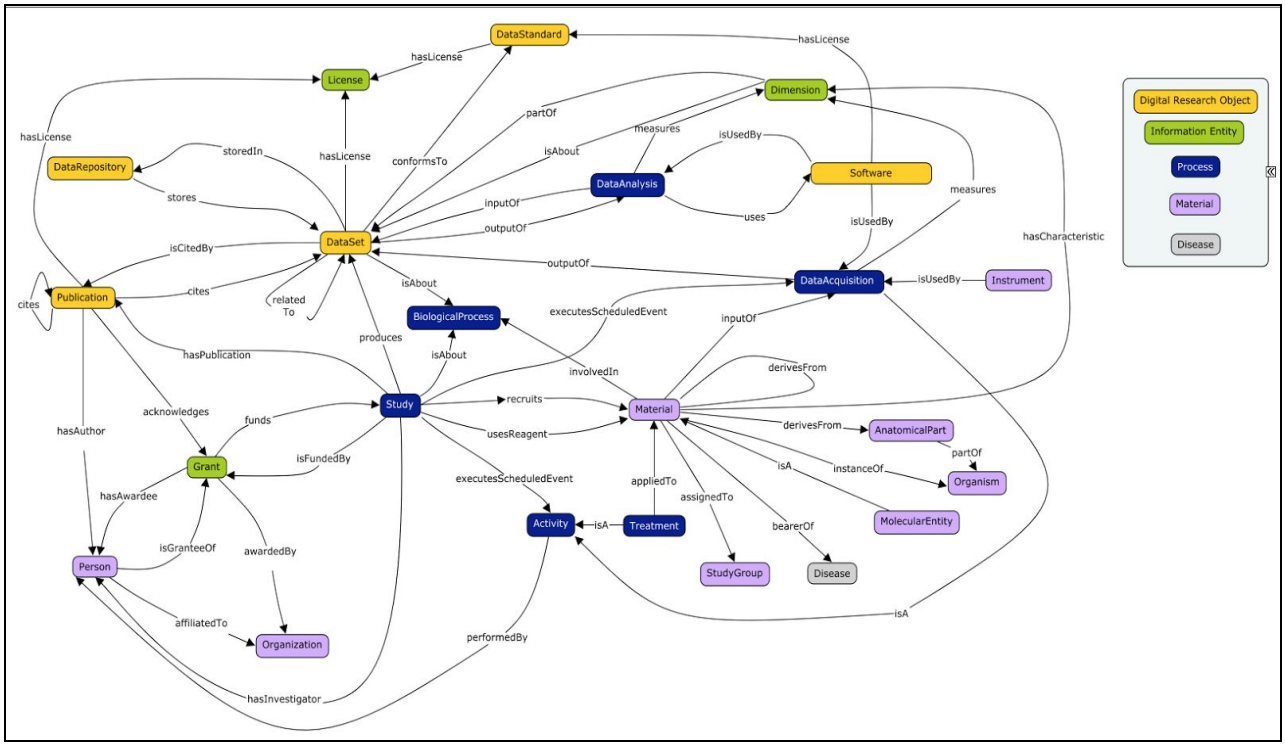
Authors: Alejandra Gonzalez-Beltran[1], Philippe Rocca-Serra[1], Susanna-Assunta Sansone[1] and WG3 members.

This section describes the initial set of 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

A schematic 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, 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 Data Discovery Index prototype should retrieve information from. This initial model will be implemented by the bioCADDIE Core Development Team and tested with a variety of sources. Subsequently, it will be iteratively reviewed (modified, simplified and/or enriched) as needed, as part of the phase 2 activities of the WG3.

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



3.2. JSON Schemas

The metadata elements are also available as machine readable JSON schemata from bioCADDIE Github repository, along with example JSON files.

3.3. Detailed Description

A full description of the metadata elements, grouped by types and color coded (to match Figure 1) in provided in Table 1, which is also available as a separate: Appendix II - NIH BD2K BioCADDIE WG3 Metadata Elements File v1 (section 5); 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); others are detailed below.

3.3.1. Cardinality and Requirement Level

Cardinality restrictions indicate the number of valid occurrences for an attribute; the RFC 2119 requirement levels indicates if a particular entity MUST/MUST NOT/SHOULD/SHOULD NOT/MAY be present, i.e. if it is compulsory, recommended, optional. 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 NIH BD2K Data Discovery Index prototype's metadata, and therefore if it has the potential to fulfil the relevant competency questions.

3.3.2. Links to Use Cases and Schemas/Models

The 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.

Table 1. A full description of the metadata elements grouped by types and color coded to match Figure 1.

ENTITY	PROPERTY	DEFINITION	TYPE	CARDINALITY	REQUIREMENT LEVEL	COMPETENCY QUESTION
Dataset		**A set of dimensions about an entity being observed. A collection of data, published or curated by a single agent, and available for access or download in one or more formats (from DCAT: http://www.w3.org/TR/vocab-dcat/#Class:_Dataset) A body of structured information describing some topic(s) of interest (from: http://schema.org/Dataset)**				**BGUC5-2;BGUC5-4;BGUC5-5;UC2;UC15;WPUC5-p7;WPUC7-p7;WPUC8-p7;WPUC10-p7**
	identifier	a code uniquely identifying the dataset	**IRI**	0..n	SHOULD	**BGUC5**
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	title	one sentence, usually short description of an entity, here the dataset	**string**	1	MUST	
	description	a textual narrative, comprised of one or more statements describing the dataset	**string**	0..1	SHOULD	
	dataType	a term, ideally from a controlled terminology, identifying the nature of the data, placing it in a typology	**string or IRI**	1..n	MUST	**BGUC1-1;BGUC1-2;BGUC3-2;BGUC3-3;BGUC5;BGUC5-1;WPUC1;WPUC2;WPUC3;WPUC9-p7;UC1**
	conformsTo	a property used to indicate if the dataset meets the requirements and constraints	**DataStandard**	0..n	MAY	**BGUC5-7;WPUC9-p7**

		defined by specific community data standard				
	storedIn	a property used to specify the data repository(ies) hosting the dataset	**DataRepository**	0..n	MAY	**BGUC1-1;UC2**
	hasPartDimension	a property used to identify, specify and list the different dimensions (granular components) making up a dataset.	**Dimension**	0..n	MAY	**BGUC2;BGUC5-4**
	isCitedBy	a property used to specify links to relevant publication(s)	**Publication**	0..n	MAY	**BGUC5-2**
	license	a property used to specify relevant terms of usage and license	**License**	0..n	SHOULD	**BGUC5-1;BGUC5-4; BGUC5-8**
	downloadURL	a property used to specify a URL where the dataset can be obtained	**URL**	0..n	SHOULD	
	producedBy	a property used to associate a study process which generated a given dataset, if any	**Study**	0..1	SHOULD	
	creator	a property to list the person(s) or organization(s) which contributed to the creation of the dataset.	**Person or Organization**	1..n	MUST	**UC2**
	date	a property used to specify any date relevant to the dataset	**date**	1..n	MUST	
	dateType	a property qualifier to the 'data' attribute. The type of date, used to specify the process which is being timestamped by the date attribute value, ideally comes from a controlled terminology.	**string or IRI**	1, if date is available	(MUST)	
	version	an optional property used to specify a release point for the dataset when applicable	**string**	0..1	SHOULD	**WPUC5-p7**
	isAboutBiological Process	a property used to specify the biological processes relevant to the study, ideally from a controlled vocabulary/ontology	**Biological Process**	0..n	SHOULD	
	size	a property used to specify the size of the dataset	**number**	0..1	MAY	**BGUC5-1**
	relatedDataset	a property to indicate if the dataset relies on other dataset(s)	**Dataset**	0..n	MAY	
DataStandard		**A format, reporting guideline, terminology. It is**				**BGUC5-7;UC15;WPUC9-p7**

		used to indicate whether the dataset conforms to a particular community norm or specification.				
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	the name of the standard	**string**	1	MUST	
	homepage	URL of the homepage of the standard	**URI**	1	MUST	
	dataStandardType	a descriptor (ideally from a controlled vocabulary) providing information about the nature of the information resource	**string or IRI**	1	MUST	**WPUC9-p7**
	license	a property used to specify to the terms of use of the standard	**License**	0..n	SHOULD	**BGUC5-4**
	version	an optional property used to specify a release point for the repository when applicable	**string**	0..1	SHOULD	
DataRepository		**A repository or catalog of dataset**				**BGUC1-1;UC2;UC15**
	identifier	a code uniquely identifying the data repository	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	the name of the data repository	**string**	1	MUST	**BGUC1-1;UC2**
	homepage	URL of the homepage of the data repository	**URI**	1	MUST	
	dataRepositoryDataType	a descriptor (ideally from a controlled vocabulary) providing information about the nature of the datasets in the repository	**string or IRI**	0..n	1..n	**SPUC1;SPUC7-2**
	license	a property used to specify to the terms of use of the standard	**License**	0..n	SHOULD	**BGUC5-4**

	version	an optional property used to specify a release point for the repository when applicable	**string**	0..1	SHOULD	
	creator	a property used to specify the person(s) or organization(s) responsible for the repository	**Person or Organization**	1..n	MUST	
Software		**A digital entity containing sets of instructions and operation, which allows computation and operation of and by computer**				**SPUC11;SPUC10**
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1	(MUST)	
	name	a property to specify the name of the software program	**string**	1	MUST	
	license	a property used to specify to the terms of use of the software	**License**	0..n	SHOULD	
	version	an optional property used to specify a release point for the software when applicable	**string**	0..1	SHOULD	
Publication		**A digital document made available by a publisher.**				**BGUC5-2;WPUC5-p7;WPUC10-p7;UC2**
	identifier	a code uniquely identifying an entity locally to a system or globally.	**IRI**	0..n	SHOULD	**BGUC5-6;BGUC5-2;SPUC14;SPUC5;UC2**
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	publicationType	type of publication type, delegated to an external vocabulary/resource	**string or IRI**	0..1	SHOULD	
	title	a property to specify the title of the publication	**string**	1	MUST	
	date	a property to specify the date of the publication	**date-time**	1	MUST	
	author		**Person or Organization**	1..n	MUST	**BGUC5-6**

		a property to specify the person(s) responsible for the publication				
	acknowledges	a property to specify the grant(s) which funded and supported the work reported by the publication	**Grant**	0..n	SHOULD	
	cites	a property to specify a reference to a dataset or publication	**Dataset or Publication**	0..n	MAY	**BGUC5-2**
	license	a property used to specify to the terms of use of the publication	**License**	0..n	SHOULD	
Grant		**An allocated sum of funds given by a government or other organization for a particular purpose**				**BGUC5-6**
	identifier	a code uniquely identifying an entity locally to a system or globally.	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	a property to specify the name of the grant and its funding program	**string**	1	MUST	
	funds	a property to specify the study (or dataset) supported by the grant	**Study (or Dataset)**	0..n	SHOULD	
	funder	a property to specify the organization(s) which has awarded the funds supporting the project	**(Person or Organization) and role funder**	1..n	MUST	**BGUC5-6;WPUC7-p7;WPUC8-p7;WPUC10-p7;UC1**
	awardee	a property to specify the person(s) or organization(s) which received the funds supporting the project	**Person or Organization**	0..n	SHOULD	
License		**A legal document giving official permission to do something with a Resource. It is assumed that an external vocabulary will describe with sufficient granularity the permission for redistribution, modification, derivation,**				**BGUC5-4,BGUC5-8**

		reuse, etc. and conditions for citation/acknowledgment.				
	identifier	a code uniquely identifying an entity locally to a system or globally.	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	a property to specify the name of the license	**string**	1	MUST	
	version	an optional property used to specify the version of the license	**string**	0..1	SHOULD	
	creator	a property to specify the person(s) or organization(s) responsible for writing the license	**Person or Organization**	0..n	SHOULD	
Dimension		**A feature of an entity, i.e. an individual measurable property (both quantitative or qualitative) of the entity being observed**				**BGUC2;BGUC4;BGUC5-1;BGUC5-4;PB1**
	identifier	a code uniquely identifying the dimension	**IRI**	0..1	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	the name of the dimension measured or observed during the data acquisition process (e.g. signal intensity, standard deviation), ideally from a controlled terminology	**string or IRI**	1	MUST	**BGUC5-10,WPUC3, SPUC6,SPUC1**
	dimensionType	a term, ideally from a controlled terminology, identifying the nature of the dimension, placing it in a typology	**string or IRI**	1..n	MUST	
	values	a property used to specify the actual collections of values collected for that dimension	**set**	0..n	SHOULD	**BGUC2**
	unit	an optional property used to specify a reference measurement unit associated with scalar dimensions. Ideally,	**string or IRI**	0..1	MAY	

		unit should be coming from a reference controlled terminology.				
	isAbout	an optional property used to specify what the dimension is about, it could be about material (the heights of the patients) or about a dataset (the standard deviation or the set of outliers or a quality indicator of a dataset)	**Dataset or Material**	0..n	MAY	**BGUC5-4;WPUC9-p7 ;PB1**
	partOf	a property used to specify the dataset(s) this dimension belongs to	**Dataset**	1..n	MUST	
Activity	**Superclass for Study, DataAnalysis, DataAcquisitio n**	**A type of process scheduled in a study**				**BGUC5-5;BGUC5-7; BGUC5-6**
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierSche me	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	title	a property to specify the title/name of the process	**string**	1	MUST	
	startDate	a timestamp to record the starting point of the process	**date**	0..1	SHOULD	
	endDate	a timestamp to record the ending point of the process	**date**	0..1	SHOULD	
	duration	a property to specify the duration of the activity	**string**	0..1	MAY	
	location	a property to specify where the process is performed	**string or IRI**	0..1	SHOULD	
	performedBy	a property to specify the person(s) or organisation(s) responsible for executing the process	**Person or Organization**	0..n	SHOULD	**BGUC5-7**
Study	**(subclassOf Activity)**	**Process to acquire data on a sample and attempt to draw conclusions about the population the sample has been selected from, executing a plan and design**				**BGUC4;BGUC5-3**

	identifier	a code uniquely identifying an entity locally to a system or globally.	**IRI**	0..n	SHOULD	**SPUC8; BGUC5-3; BGUC4;SPUC15;SPUC19**
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	**BGUC1-2**
	title	one sentence, usually short description of an entity, here the study	**string**	1	MUST	
	schedulesActivity	a property used to specify the activities scheduled by the study	**Activity**	0..n	SHOULD	
	schedulesDataAcquisition	a property used to specific the kind of techniques and response variables used during a study for acquiring signal	**DataAcquisition**	1..n	MUST	
	studyType	a property used to specify the type of study, e.g. intervention or observation or meta-analysis	**string or IRI**	1	MUST	**BGUC1-2;SPUC19**
	hasStudyGroup	a property used to list the different study groups associated with a study	**StudyGroup**	0..n	MAY	
	recruits	a property used to specify which materials are the subjects and participants of the study	**Material**	0..n	MAY	**BGUC4**
	usesReagent	a property used to specify which materials are used as reagents (but not subjects) of the study	**Material**	0..n	MAY	
	isAboutBiological Process	a property used to specify the biological processes relevant to the study (ideally from a controlled vocabulary/ontology)	**Biological Process**	0..n	SHOULD	**BGUC2;BGUC4**
	startDate	a timestamp to record the starting point of the process	**date**	0..1	SHOULD	
	endDate	a timestamp to record the ending point of the process	**date**	0..1	SHOULD	
	duration	a property to specify the duration of the activity	**string**	0..1	MAY	
	location	a property to specify where the process is performed.	**string or IRI**	0..1	SHOULD	
	performedBy	a property to specify the person(s) or organisation(s) responsible for executing the process	**Person or Organization**	0..n	SHOULD	

	keywords	a property used to provide a list of terms providing insights into the main topic and feature of the study	**string or IRI**	0..n	MAY	
	resultsIn	a property used to specify the collection of data during assays over the course of a study	**Dataset**	1..n	MUST	
Treatment	**(subclassOf Activity)**	**Process, part of a study, consisting in exposing participants to the study to different conditions or group those participants into different categories based on specific criteria and compare their outcomes**				**BGUC3-4;BGUC3-5; BGUC4;BGUC5-3;BGUC5-11;SPUC8;SPUC9;BGUC5-10**
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	agent	a property used to specify the nature of the perturbation or intervention used in the study	**MolecularEntity;Material;Activity;IRI;string**	0..1	SHOULD	**BGUC3-1;BGUC3-3; BGUC5-3**
	intensity	a property used to specify how acute the perturbation is	**string if qualitative, double if quantitative**	0..1	SHOULD	
	title	one sentence, usually short description of the study	**string**	1	MUST	
	startDate	a timestamp to record the starting point of the process	**date**	0..1	SHOULD	
	endDate	a timestamp to record the ending point of the process	**date**	0..1	SHOULD	
	duration	a property to specify the duration of the activity	**string**	0..1	MAY	
	location	a property to specify where the process is performed	**string or IRI**	0..1	SHOULD	
	performedBy	a property to specify the person(s) or organisation(s) responsible for executing the process	**Person or Organization**	0..n	SHOULD	
	concomitance	a property used to specify if more that one perturbations are applied at the same time to the same subject	**boolean**	0..1	MAY	

	order	a property used to specify the rank in which perturbations are being applied to study subjects	**integer**	0..1	MAY	
DataAcquisition	**(subclassOf Activity)**	**Process of generating data through measurement made with specific techniques**				
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	title	a property to specify the name of the technology used to acquire signal	**string**	1	MUST	
	uses	a property to specify the instrument(s) used to acquire the signal	**Instrument or Software**	0..n	MAY	
	startDate	a timestamp to record the starting point of the process	**date**	0..1	SHOULD	
	endDate	a timestamp to record the ending point of the process	**date**	0..1	SHOULD	
	duration	a property to specify the duration of the activity	**string**	0..1	MAY	
	location	a property to specify where the process is performed	**string or IRI**	0..1	SHOULD	
	performedBy	a property to specify the person(s) or organisation(s) responsible for executing the process	**Person or Organisation**	0..n	SHOULD	
	measures	a property to specify the dimension(s) being acquired as signal	**Dimension**	1..n	MUST	**BGUC2**
DataAnalysis	**(subclassOf Activity)**	**Process of transforming data and producing data**				**SPUC4,SPUC16,WPUC6-p7,BGUC5-9,BGUC5-11;WPUC9-p7**
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	

	title	a property to specify the name of the analysis being performed	**string**	1	MUST	
	startDate	a timestamp to record the starting point of the process	**date**	0..1	SHOULD	
	endDate	a timestamp to record the ending point of the process	**date**	0..1	SHOULD	
	duration	a property to specify the duration of the activity	**string**	0..1	MAY	
	location	a property to specify where the process is performed	**string or IRI**	0..1	SHOULD	
	uses	a property to specify any software package that may have been using in the data analysis process	**Software**	0..n	MAY	
	input	a property to specify the dataset used as input	**Dataset**	1..n	MUST	
	output	a property to specify the dataset resulting from applying the technology	**Dataset**	1..n	MUST	
Biological Process		**A biological process is a recognized series of events or molecular functions (from: http://geneontology.org/page/biological-process-ontology-guidelines)**				**BGUC2;BGUC3-1;BGUC3-2;BGUC3-4;BGUC5-11;SPUC13;WPUC2;WPUC3;WPUC4**
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	a property to specify the name of the biological process	**string**	1	MUST	**BGUC2**
Material		**A physical entity, part of collection or used in a study**				**BGUC3-3;BGUC3-5;BGUC5;BGUC5-1;BGUC5-9;BGUC5-11;PB1;SPUC13;WPUC6-p7**
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent	**string**	1, if identifier is available	(MUST)	

		information about how identifiers are formed, maintained and minted				
	name	a property to specify the name of the material	**string**	1	MUST	
	derivesFrom	a property to indicate that this material originated from another material	**Material**	0..n	MAY	
	anatomicalPart	a property to indicate that this material originated from an anatomical part (ideally from a controlled vocabulary/ontology)	**AnatomicalPart**	0..n	MAY	
	bearerOfDisease	a property to specify the pathology affecting the material used in the study or refered to in the dataset (ideally from a controlled vocabulary/ontology)	**Disease**	0..n	MAY	**BGUC1-1;BGUC1-2;BGUC1-3;BGUC5,BGUC5-4,BGUC5-6,BGUC5-8,BGUC-5-9,SPUC7-3,WPUC1**
	instanceOfOrganism	a property to specify the organism this is an instance of - when the whole organism is considered and participates in the study (ideally from a controlled vocabulary/ontology)	**Organism**	0..n	MAY	**BGUC2**
	derivesFromOrganism	a property to specify the organism this material derives from - for instance when samples are taken from an organism (ideally from a controlled vocabulary/ontology)	**Organism**	0..n	MAY	**BGUC2**
	involvedInBiologicalProcess	a property to specify that the material is involved in a particular biological process (ideally from a controlled vocabulary/ontology)	**Biological Process**	0..n	MAY	**BGUC2;BGUC3-1;BGUC3-2;BGUC4;SPUC18**
	inputOf	a property to specify the pathology affected the material used in the study or refered to in the dataset	**DataAcquisition**	0..n	MAY	
	hasCharacteristic	a property to specify the characteristic information denoting the material	**Dimension**	0..n	MAY	**BGUC2**
	role	a property to specify the role played by a material in a study	**string or IRI**	0..n	SHOULD	
StudyGroup		**A collection of entities known as study subjects based on a set of specified criteria and rules; synonyms: population, cohort**				

	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	a property to specify the name of the study group	**string**	1	MUST	
	size	a property to specify the number of members in the group	**integer**	0..1	MAY	
	member	a property to list references entities of type material make up a group	**Material**	1..n	MUST	
	selectionCriteria	a property to list the attributes of material and the values those should meet in order to be part of the group	**string or IRI**	0..n	SHOULD	**BGUC5-10;BGUC5-9;BGUC3-5;BGUC5-11;SPUC1;SPUC3;PRE3**
MolecularEntity	**(subclass of Material)**	**A physical entity of molecular scales such as proteins, nucleic acids, chemical materials. They can be abiotic, biological or synthetic origin**				**BGUC1-3;SPUC18;WPUC6-p7**
	identifier	a code uniquely identifying an entity locally to a system or globally.	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1	(MUST)	
	name	a property to specify the name of the molecular entity	**string**	1	MUST	**BGUC1-3;BGUC3-1;BGUC3-2;BGUC3-3;BGUC5-3;BGUC5-11**
	structure	a property to specify the primary sequence of the molecular entity	**string**	0..n	MAY	
	role	a property to specify the role played by a molecular entity in a study	**string or IRI**	0..n	SHOULD	
Person		**A human being**				**UC2**
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	

	identifierScheme	If an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	firstName	a string denoting the given name of a person	**string**	1	MUST	
	lastName	a string denoting the family name of a person	**string**	1	MUST	
	email	a string denoting an SMTP electronic mail address, contains @ character	**email**	0..n	SHOULD	
	affiliation	a string or crossreference denoting a legal entity or physical entity corresponding to a business or administration	**Organization**	0..n	SHOULD	
	role	roles assumed by a person, defined in an external resource	**string or IRI**	0..n	SHOULD	**(has_role author) BGUC5-6, UC2**
Organization		**Legal or physical entity corresponding to a business or administration**				
	identifier	a code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	If an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	a property to specify the name of the organization	**string**	1	MUST	
	abbreviation	a property to specify the shortname, abbreviation associated to the organization	**string**	0..1	MAY	
	postalAddress	a property to specify the postal, street address associated to an organization	**string**	0..1	MAY	
	role	A property that identifies the role of the organization (ideally from a controlled vocabulary/ontology)	**string or IRI**	0..n	SHOULD	**UC1; SPUC5**
Instrument		**A physical entity produced by humans to perform specific tasks or functions**				
	identifier	A code uniquely identifying an entity locally to a system or globally	**IRI**	0..n	SHOULD	

	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	a property to specify the name of the instrument	**string**	1	MUST	**WPUC4;SPUC10**
	instrumentType	a property to specify the type of the instrument/technology	**string or IRI**	0..1	SHOULD	**WPUC4;SPUC10**
	isUsedBy	a property to specify the activity that makes use of this instrument	**DataAcquisition**	0..n	MAY	
	manufacturer	a property to specify the organisation which produced the instrument	**Person or Organization**	0..n	MAY	
Organism		**A living entity**				**BGUC2**
	identifier	a code uniquely identifying a taxonomic information entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	a property to specify the scientific name of an organism	**string**	1	MUST	
AnatomicalPart		**A structure that is part of a multicellular organism**				
	identifier	a code uniquely identifying an anatomical entity locally to a system or globally	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	a property to specify the name of the anatomical entity or anatomical part	**string**	1	MUST	
Disease		**A disorder of structure or function in a human, animal, or plant, that produces specific symptoms or that affects a specific location.**				**BGUC1-1,BGUC1-2,BGUC1-3,BGUC5,BGUC5-4,BGUC5-6,BGUC5-8,BGUC-5-9,SPUC7-3,WPUC1**

	identifier	a code uniquely identifying an entity locally to a system or globally.	**IRI**	0..n	SHOULD	
	identifierScheme	if an identifier is supplied, the identifier scheme represent information about how identifiers are formed, maintained and minted	**string**	1, if identifier is available	(MUST)	
	name	a property to specify the name of the disease	**string**	1	MUST	

4. Appendix I

Authors: Alejandra Gonzalez-Beltran[1], Philippe Rocca-Serra[1] and WG3 members.

NIH BD2K bioCADDIE WG3 Metadata Mapping File v1, also downloadable from the bioCADDIE Github repository. 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], Philippe Rocca-Serra[1], Susanna-Assunta Sansone[1] and WG3 members.

NIH BD2K bioCADDIE WG3 Metadata Elements File v1, also downloadable from the bioCADDIE Github repository. The file describes the metadata elements, grouped by types, providing details on their: definition, attributes, requirements, cardinality and requirement level.