dats-doc Documentation

Release 0.1

Philippe Rocca-Serra and Alejandra Gonzalez-Beltran

Aug 30, 2018

Contents

NOTE: this documentation has been replaced by the documentation at the Data Tag Suite github organization

CHAPTER 1

Introduction:

DATS, which stands for DAta Tag Suite, is a data description model designed and produced to describe datasets being ingested in DataMed, a prototype for data discovery developed as part of the NIH Big Data 2 Knowledge bioCADDIE project.

For more information about DATS, please check the DATS pre-print available in bioarxiv. For more information about DataMed, please check the DataMed pre-print available in bioarxiv. For more information about the objectives of the bioCADDIE project, please have a look at the bioCADDIE White Paper.

This documentation describes the DATS model and how to use it. More details about how DATS was designed and how it relates to other models can be found in the aforementioned documents as well as in the documents accompanying each of the releases.

Table of Contents:

1.1 First Steps with DATS

This document offers an overview of the DATS model from a practical perspective, detailing how DATS may be used to document a specific dataset.

The DATS model is centered around the Dataset entity, which supports most of the relevant information about the data being observed.

The main building blocks of the DATS model are defined as "entities", which and for convenience purposes, may be compared to the different "sections" of information in a flat document. Each entity has a number of properties that are instantiated either as other entities or as direct entries. For the latter, information may may be structured (e.g., integer, date, URI) or unstructured (string, or free text entries).

First and foremost, *Dataset* entity aims to cater for essential provenance information: who, when, what, why, where, and how. By answering these questions, each dataset source will define its own view on what a dataset is. The *Dataset* entity is also designed to declare which variables were measured and what type of data was collected.

1.1.1 *What* is the dataset about?

The nature of the information available in a dataset can be recorded via the DATS Dimension entity. It is the object to use for reporting variables measured and for which data have been collected.

The DATS Dimension object can be qualitied using the DATS DataType entity.

The DATS *DataType* covers four aspects of a variable's nature: type of information (what the data is about), method (how the data was generated), platform (the instrumentation, software and reagents used to generate the data), and instrument (the specific device used to generate the data).

Importantly, it is key to remember that Dataset may be constitutive parts of another Dataset. Each of these dataset parts can be used to describe a particular aspect of a dataset in greater details. For instance, a dataset describing a multi-omics experiment may contain several datasets, one focusing on transcriptomics, one focusing on metabolomics and so on.

1.1.2 *Why* was the data produced?

As a *Dataset* property, the "description" is a textual narrative that typically indicates the dataset's purpose and why it was produced.

In addition, in the extended DATS it is possible to describe the *Study* that produced one, or several related datasets, including the purpose, objective, or hypothesis that gave origin to the dataset(s) defined as belonging to a study.

Related studies may also be grouped to constitute a series.

Tracking dataset spatial and temporal properties

1.1.3 *Where* was the dataset collected and where was it produced?

The DATS Dataset property *spatialCoverage* includes a description of the geography covered by the dataset and/or measured by the dataset's dimensions or variables.

spatialCoverage is instantiated within a Place entity, which maps to the entity bearing the same name in schema.org (http://schema.org/Place), to "geoLocation" in the DataCite schema (http://schema.datacite.org/meta/kernel-4.0/) and to "Feature" in GeoJSON (https://tools.ietf.org/html/rfc7946).

1.1.4 *When* was the dataset produced?

DATS model provides a Date object to records key *Date(s)* associated with the description of a *Dataset*.

For each *Date*, users have to identify its type, in relation to a specific event (e.g. creation, update, validation, verification, deprecation. . .).

Such generic mechanism of providing *Date* and temporal information offers flexibility and extensibility. Dates may be repeated and differentiated by type. This allows for extensions to new types of dates that may be required in specific scenarios. The actual definition of the types is delegated to existing ontologies.

1.1.5 *Who* produced the dataset?

Using the Dataset's "creators" property, DATS records the Person and/or Organization associated with the dataset, and supports documenting their roles (e.g., creator, curator, developer, funder, principal investigator).

1.1.6 *Where* and *How* can the dataset be accessed?

DATS provides for a comprehensive description of the ways to access a Dataset. This information can be reported in the Access entity, that is part of DatasetDistribution as well as part of the description of a DataRepository. It covers information such as the dataset landing page and/or access URL if available, a description of the type of access (such as download, remote access, remote service, enclave or not available) as well as any authorization or authentication needed to access the dataset.

1.2 DATS Model

Table 1: DATS specifications

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
`dataset`	`identifier`	Primary identifiers for the dataset.	IdentifiersInformation	0..n	SHOULD	BGUC5	
	`relatedIdentifiers`	Related identifiers for the dataset.	IdentifiersInformation	0..n	SHOULD	BGUC5	
	`alternateIdentifiers`	Alternate identifiers for the dataset.	AlternateIdentifiersInformation	0..n	MAY		
	`title`	The name of the dataset, usually one sentence or short description of the dataset.	string	1	MUST	BGUC5	DataCite[/resource/titles];Dat //schema.org/headline{]}; HCLS{[}(dct:title,rdf:langString)]
	`types`	A term, ideally from a controlled terminology, identifying the dataset type or nature of the data, placing it in a typology.	DataType	`1..n`	`MUST`	BGUC1-1;BGUC1-2;BGUC3-2;BGUC3-3;BGUC5;BGUC5-1;WPUC1;WPUC2;WPUC3;WPUC9-p7;UC1	For example: microscopy imaging, gene expression profile, genomic sequence, fMRI, pathway simulation.
	`creators`	The person(s) or organization(s) which contributed to the creation of the dataset.	Person or Organization	1..n	MUST	UC2	

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	`dates`	Relevant dates for the dataset, a date must be added, e.g. creation date or last modification date should be added.	Date	0..n	MAY		
	`distributions`	The distribution(s) by which datasets are made available (for example: mySQL dump).	DataSet Distribution	`0..n`	`SHOULD`		
	`dimensions`	The different dimensions (granular components) making up a dataset.	Dimension	`0..n`	`MAY`	BGUC2;BGUC5-4	
	`isCitedBy`	The relevant publication(s) describing how the dataset was produced or used.	Publication	`0..n`	`MAY`	BGUC5-2	
	`producedBy`	A study process which generated a given dataset, if any.	Study	`0..1`	`SHOULD`		
	`isAbout`	Different entiies (biological entity, taxonomic information, disease, molecular entity, anatomical part, treatment) associated with this dataset.	BiologicalEntity or TaxonomicInformation or Disease or MolecularEntity or AnatomicalPart or Treatment	`0..n`	`SHOULD`		
	`hasPart`	A Dataset that is a subset of this Dataset; Datasets declaring the 'hasPart' relationship are considered a collection of Datasets, the aggregation criteria could be included in the 'description' field.	Dataset	`0..n`	`MAY`		
	`keywords`	Tags associated with the dataset, which will help in its discovery.	Annotation	`0..n`	`MAY`		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	acknowledges	The grant(s) which funded and supported the work reported by the dataset.	Grant	0..n	MAY		
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
DatasetDistribution		"A specific available form of a dataset. Each dataset might be available in different forms, these forms might represent different formats of the dataset or different endpoints. Examples of distributions include a downloadable CSV file, an API or an RSS feed. (From DCAT) "				BGUC5	
	identifiers	Primary identifiers for the dataset distribution.	IdentifiersInformation	1..n	SHOULD	BGUC5	
	alternateIdentifiers	Alternate identifiers for the dataset distribution.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the dataset distribution.	RelatedIdentifiersInformation	0..n	MAY		
	title	"The name of the dataset distribution, usually one sentece or short description of the dataset."	string	0..1	MAY		
	description	An textual narrative comprised of one or more statements describing the dataset distribution.	string	0..1	SHOULD		
	dates	"Relevant dates for the datasets, a date must be added, e.g. creation date or last modification date should be added."	Date	1..n	MUST		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	‘‘storedIn ‘‘	The data repository(ies) hosting the dataset.	DataRepository	0..n	MAY	BGUC1-1;UC2	“While from the DDI perspective, every dataset may be coming from a data repository, we put a less strict requirement allowing for datasets available online and not in a repository.”
	version	A release point for the dataset when applicable.	string	0..1	SHOULD	WPUC5-p7	
	accessModalities	The information about access modality for the dataset.	Access	1..n	MUST		
	licenses	The terms of use of the data standard.	License	0..n	SHOULD	BGUC5-4	

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	curationStatus	The level of curation of the dataset distribution.	Annotation	0..n	MAY		"E.g. manually or authomatic or both, other values such as https://wiki.nci.nih.gov/display/CTRPdoc/Curation+Status+Definitions+-+Include+v4.3.1"
	conformsTo	A data standard whose requirements and constraints are met by the dataset.	DataStandard	0..n	MAY	BGUC5-7;WPUC9-p7	
	format	The technical format of the dataset distribution. Use the file extension or MIME type when possible. (Definition adapted from DataCite)	string	0..n	MAY		"e.g. PDF, XML, MPG or application/pdf, text/xml, video/mpeg"

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	qualifiers	"One or more characteristics of the dataset distribution (e.g. how it relates to other distributions, if the data is raw or processed, compressed or encrypted). "	Annotation or CategoryValuesPair	0..n	MAY		"e.g. indicate if the distribution is isomorphic (corresponds completely with the dataset), a derivative from the dataset, or is a partial distribution of the dataset. These qualifiers can also indicate if the distribution refers to raw, processed or summarised data. It could also refer to the data being encrypted or compressed."
	"size "	The size of the dataset.	number	0..1	MAY	BGUC5-1	

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	unit	“The unit of measurement used to estimate the size of the dataset (e.g, petabyte). Ideally, the unit should be coming from a reference controlled terminology.”	Annotation	“1, if size is reported”	(MUST)		
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
DataStandard		“A format, reporting guideline, terminology. It is used to indicate whether the dataset conforms to a particular community norm or specification.”				BGUC5-7;UC15;WPUC9-p7	
	identifiers	Primary identifiers for the standard.	IdentifiersInformation	0..n	SHOULD	BGUC5	
	alternateIdentifiers	Alternate identifiers for the standard.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the standard.	RelatedIdentifiersInformation	0..n	MAY		
	name	“The name of the standard (e.g. FASTQ, CDISC STDM, ISO8601)”	string	1	MUST		
	type	“The nature of the information resource, ideally specified with a controlled vocabulary or ontology (.e.g model or format, vocabulary, reporting guideline).”	Annotation	1	MUST	WPUC9-p7	
	description	An textual narrative comprised of one or more statements describing the data standard.	string	0..1	SHOULD		
	licenses	The terms of use of the data standard.	License	0..n	SHOULD	BGUC5-4	
	version	A release point for the repository when applicable.	string	0..1	SHOULD		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	`extraProperties`	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
`DataRepository`		A repository or catalog of datasets. It could be a primary repository or a repository that aggregates data existing in other repositories.				BGUC1-1;UC2;UC15	
	`identifiers`	Primary identifiers for the data repository.	IdentifiersInformation	0..n	SHOULD	BGUC5	
	`alternateIdentifiers`	Alternate identifiers for the data repository.	AlternateIdentifiersInformation	0..n	MAY		
	`relatedIdentifiers`	Related identifiers for the data repository.	RelatedIdentifiersInformation	0..n	MAY		
	`name`	The name of the data repository.	string	1	MUST	BGUC1-1;UC2	
	`description`	An textual narrative comprised of one or more statements describing the data repository.	string	0..1	SHOULD		
	`dates`	Relevant dates for the data repository.	Date	0..n	MAY		
	`scopes`	"Information about the nature of the datasets in the repository, ideally from a controlled vocabulary or ontology (e.g. transcription profile, sequence reads, molecular structure, image, DNA sequence, NMR spectra)."	Annotation	0..n	1..n	SPUC1;SPUC7-2	
	`types`	"A descriptor (ideally from a controlled vocabulary) providing information about the type of repository, such as primary resource or aggregator."	Annotation	0..n	SHOULD		
	`licenses`	The terms of use of the data repository.	License	0..n	SHOULD	BGUC5-4	

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	version	"A release point for the repository, when applicable."	string	0..1	SHOULD		
	publishers	The person(s) or organization(s) responsible for the repository and its availability.	Person or Organization	0..n	SHOULD		
	aggregatorOf	The DataRepositories aggregated by this repository. This property will be empty for primary repositories.	DataRepository	0..n	MAY		
	accessModalities	The information about access modality for the data repository.	Access	1..n	MAY		
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
Software		"A digital entity containing sets of instructions and operation, which allows computation and operation of and by computer."				SPUC11;SPUC10	
	identifiers	Primary identifiers for the software.	IdentifiersInformation	0..n	SHOULD	BGUC5	
	alternateIdentifiers	Alternate identifiers for the software.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the software.	RelatedIdentifiersInformation	0..n	MAY		
	name	The name of the software.	string	1	MUST		
	licenses	The terms of use of the software.	License	0..n	SHOULD		
	isUsedBy	The data acquisition activity that makes use of this software.	DataAcquisition or DataAnalysis	0..n	MAY		
	manufacturer	The person or organisation that produced the software.	Person or Organization	0..1	MAY		e.g. Adobe
	version	A release point for the software.	string	0..1	SHOULD		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
Publication		A (digital) document made available by a publisher.				BGUC5-2;WPUC5-p7;WPUC10-p7;UC2	
	identifiers	Primary identifiers for the publication.	IdentifiersInformation	1..n	SHOULD	BGUC5	
	alternateIdentifiers	Alternate identifiers for the publication.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the publication.	RelatedIdentifiersInformation	0..n	MAY		
	title	“The name of the publication, usually one sentece or short description of the publication.”	string	1	SHOULD		
	““dates ‘‘	“Relevant dates, the date of the publication must be provided. “	Date	1..n	SHOULD		
	type	“Publication type, ideally delegated to an external vocabulary/resource.”	Annotation	0..1	SHOULD		“e.g. book, article, weblog, chapter, review, correspondence”
	publicationVenue	The name of the publication venue where the document is published if applicable.	string	0..1	MAY		
	authorsList	The list of authors made available as a string (does not allow disambiguation).	string	0..1	SHOULD		
	authors	The person(s) and/or organisation(s) responsible for the publication.	Person or Organization	1..n	SHOULD	BGUC5-6	
	acknowledges	The grant(s) which funded and supported the work reported by the publication.	Grant	0..n	SHOULD		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	licenses	The terms of use of the publication.	License	0..n	SHOULD		
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
IdentifiersInformation		Information about the primary identifier.				BGUC5	
	identifier	A code uniquely identifying an entity locally to a system or globally.	string or IRI	0..n	SHOULD	BGUC5	
	identifierSource	The identifier source represents information about the organisation/namespace responsible for minting the identifiers. It must be provided if the identifier is provided.	string	"1, if identifier is available"	(MUST)		
AlternateIdentifierInformation		Information about an alternate identifier (other than the primary).				BGUC5	
	alternateIdentifier	An identifier or identifiers other than the primary Identifier applied to the resource being registered. (definition from DataCite)	string or IRI	0..n	MAY		
	alternateIdentifierSource	The identifier source represents information about the organisation/namespace responsible for minting the identifiers. It must be provided if the identifier is provided.	string	0..n	MAY		
RelatedIdentifiersInformation		Information about a related identifier.				BGUC5	
	relatedIdentifier	An identifier of a related resource.	string or IRI		MUST		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	relatedIdentifierSource	The identifier source represents information about the organisation/namespace responsible for minting the identifiers. It must be provided if the identifier is provided.	string		(MUST)		
	relationType	The type of the relationship corresponding to this identifier.	string or IRI		SHOULD		
Annotation		“A pair of value (string or numeric) with a corresponding ontology term (IRI), if applicable.”				BGUC5	
	“value ‘‘	A label or value (string or numeric) that might be associated with an ontology term.	string or number	1	MUST		
	ontologyTermIRI / suggested renaming = ValueIRI	The IRI of an ontology term that corresponds to value.	IRI	0..1	MAY		
Date		“Information about a calendar date or timestamp indicating day, month, year and time of an event.”				BGUC5	
	date	A date following the ISO8601 standard.	date	1	MUST		“The type of date is specified in the dateType field, following the DataCite practice. (change cardinality from 1..n to 1)”

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
Access		Information about resources that provide the means to obtain an asset (a dataset or other research object).		Description of the access conditions for the object		BGUC5	
	identifiers	Primary identifiers for the access information.	IdentifiersInformation	1..n	SHOULD		
	alternateIdentifiers	Alternate identifiers for the access information.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the access information.	RelatedIdentifiersInformation	0..n	MAY		
	landingPage	A web page that contains information about the associated dataset or other research object and a direct link to the object itself.	IRI	1	MUST		
	accessURL	"A URL from which the resource (dataset or other research object) can be retrieved, i.e. a direct link to the object itself."	IRI	0..1	SHOULD		
	types	"Method to obtain the resource, ideally specified from a controlled vocabulary or ontology."	Annotation (see worksheet 'Access Types' for CV defined by WG7)	0..n	SHOULD		"download, remote access, remote service, enclave, not available"
	authorizations	Types of verification that accessing the resource is allowed. Authorization occurs before successful authentication and refers to the process of obtaining approval to use a data set. Ideally specified from a controlled vocabulary or ontology.	Annotation (see worksheet 'Access Types' for CV defined by WG7)	0..n	SHOULD		"none, click license, registration, dual individual, dual institution"

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	authentications	“Types of verification of the credentials for accessing the resource, it is the identification process at the time of access. ideally specified from a controlled vocabulary or ontology.”	Annotation (see worksheet ‘Access Types’ for CV defined by WG7)	0..n	SHOULD		“none, simple login, multiple login”
	licenses	Terms of usage as specified on a license or data use agreement.	License	0..n	MAY	BGUC5-1;BGUC5-4;BGUC5-8	
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
Grant		An allocated sum of funds given by a government or other organization for a particular purpose				BGUC5-6	
	identifiers	Primary identifiers for the grant.	IdentifiersInformation	1..n	SHOULD	BGUC5	(change to MUST?)
	alternateIdentifiers	Alternate identifiers for the grant.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the grant.	RelatedIdentifiersInformation	0..n	MAY		
	name	The name of the grant and its funding program.	string	1	MUST		
	funds	The study or dataset supported by the grant.	Study or Dataset	0..n	SHOULD		
	funders	The person(s) or 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	
	awardees	The person(s) or organization(s) which received the funds supporting the project.	Person or Organization	0..n	SHOULD		
	extraProperties	Extra properties that do not fit in the previous specified attributes.	ExtraProperty	0..n	MAY		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
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, reuse, etc. and conditions for citation/acknowledgment.”				“BGUC5-4,BGUC5-8”	
	identifiers	Primary identifiers for the license.	IdentifiersInformation	1..n	SHOULD	BGUC5	
	alternateIdentifiers	Alternate identifiers for the license.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the license.	RelatedIdentifiersInformation	0..n	MAY		
	name	The name of the license.	string	1	MUST		
	version	The version of the license.	string	0..1	SHOULD		
	creators	The person(s) or organization(s) responsible for writing the license.	Person or Organization	0..n	SHOULD		
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
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	“e.g. demographic characteristics, quality indicator, access statistics”
	identifiers	Primary identifiers for the dimension.	IdentifiersInformation	1..n	SHOULD	BGUC5	
	alternateIdentifiers	Alternate identifiers for the dimension.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the dimension.	RelatedIdentifiersInformation	0..n	MAY		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	name	“The name of the dimension measured or observed during the data acquisition process, ideally from a controlled terminology.”	Annotation	1	MUST	“BGUC5-10,WPUC3, SPUC6,SPUC11”	“e.g. signal intensity, standard deviation”
	types	“A term, ideally from a controlled terminology, identifying the nature of the dimension, placing it in a typology.”	Annotation	1..n	MUST		“e.g. continuous, discrete, scalar, ordinal “
	partOf	The dataset(s) this dimension belongs to.	Dataset	1..n	MUST		
	description	A textual narrative comprised of one or more statements describing the dimension.	string	0..1	SHOULD		
	values	The actual collections of values collected for that dimension.	array	0..n	SHOULD	BGUC2	
	unit	“A reference measurement unit associated with scalar dimensions, ideally from a reference controlled terminology.”	Annotation	0..1	MAY		
	‘‘isAbout ‘‘	“A material or a dataset, which is the object of this dimension (this dimension is about the material - e.g. the heights of the patients - or the dataset - e.g. 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	
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	information	The measurements or facts that the data is about.	Annotation	0..1	MAY		“e.g. gene expression, protein structure, proteomics, phenotyping.”
	method	The procedure or technology used to generate the information.	Annotation	0..1	MAY		“e.g. imaging, microarray, clinical trial.”
	platform	“The set of instruments, software and reagents that are needed to generated the data.”	Annotation	0..1	MAY		“e.g. Affymetrix, NGS, mass spectrometer type”
	instrument	The specific device used to generate the data.	Annotation	0..1	MAY		
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
Material		“A physical entity, part of collection or used in a study (e.g. patient)”				BGUC3-3;BGUC3-5;BGUC5;BGUC5-1;BGUC5-9;BGUC5-11;PB1;SPUC13;WPUC6-p7	
	identifiers	Primary identifiers for the material.	IdentifiersInformation	1..n	SHOULD	BGUC5	
	alternateIdentifiers	Alternate identifiers for the material.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the material.	RelatedIdentifiersInformation	0..n	MAY		
	name	The name of the material.	string	1	MUST		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	derivesFrom	A material from which this material originated.	Material or AnatomicalPart	0..n	MAY	BGUC2	
	bearerOfDisease	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"	
	taxonomicInformation	The taxonomic information for this material (ideally specified from a controlled vocabulary/ontology).	TaxonomicInformation	0..n	MAY	BGUC2	
	involvedInBiologicalEntity	A biological process (ideally specified from a controlled vocabulary/ontology) in which the material is involved.	BiologicalEntity	0..n	MAY	BGUC2;BGUC3-1;BGUC3-2;BGUC4;SPUC18	
	characteristics	The characteristic information or attributes denoting the material.	Dimension or Material	0..n	MAY	BGUC2	
	roles	The roles played by a material.	Annotation	0..n	SHOULD		
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
Person		A human being.				UC2	
	identifiers	Primary identifiers for the person.	IdentifiersInformation	1..n	SHOULD	BGUC5	
	alternateIdentifiers	Alternate identifiers for the person.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the person.	RelatedIdentifiersInformation	0..n	MAY		
	fullName	"The first name, any middle names, and surname of a person."	string	1	SHOULD		
	firstName	The given name of the person.	string	1	MAY		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	middleInitial	The first letter of the person's middle name.	string	0..n	MAY		
	lastName	The person's family name.	string	1	SHOULD		
	email	An electronic mail address for the person.	string (format=email)	0..1	SHOULD		
	affiliations	The organizations to which the person is associated with.	Organization	0..n	SHOULD		
	roles	"The roles assumed by a person, ideally from a controlled vocabulary/ontology."	Annotation	0..n	MAY	"(has_role author) BGUC5-6, UC2"	"e.g. author, creator, contributor, awardee, submitter, researcher, patient"
	extraProperties	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		
	identifiers	Primary identifiers for the organization.	IdentifiersInformation	1..n	SHOULD	BGUC5	
	alternateIdentifiers	Alternate identifiers for the organization.	AlternateIdentifiersInformation	0..n	MAY		
	relatedIdentifiers	Related identifiers for the organization.	RelatedIdentifiersInformation	0..n	MAY		
	name	The name of the organization.	string	1	MUST		
	abbreviation	"The shortname, abbreviation associated to the organization."	string	0..1	MAY		
	postalAddress	"The postal, street address associated to the organization."	string	0..1	MAY		

Continued on next page

Table 1 – continued from previous page

Entity	Property	Definition	Value(s)	Cardinality	Requirement Level	Relevant Competency Question(s)	Notes or Example(s)
	`roles`	"The roles of the organization, ideally from a controlled vocabulary/ontology."	Annotation	0..n	MAY	UC1; SPUC5	"e.g. author, creator, contributor, awardee, submitter, researcher, patient"
	`extraProperties`	Extra properties that do not fit in the previous specified attributes.	CategoryValuesPair	0..n	MAY		

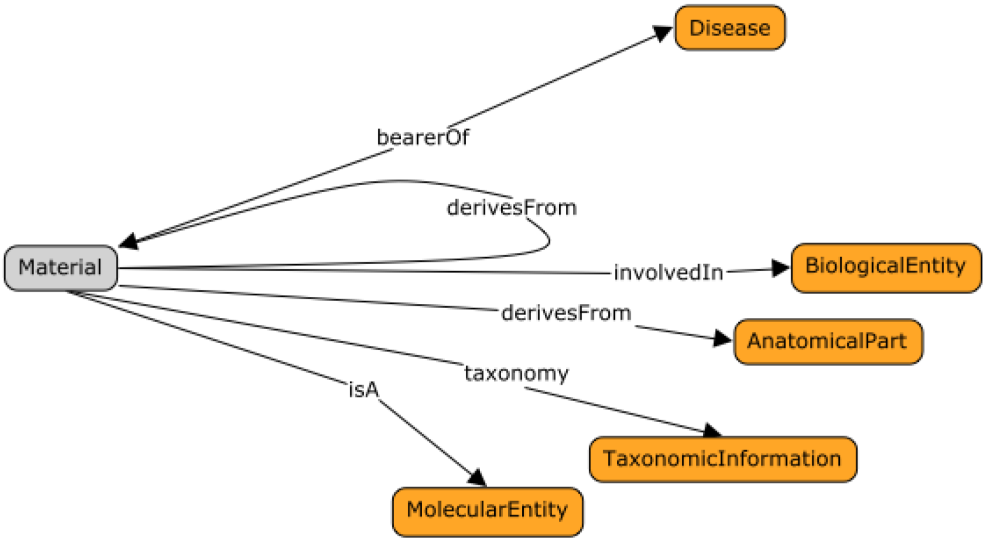
1.3 DATS Counting things:

A recurring capability query cases is that addressing the ability to assemble synthetic cohorts by interogating a collection of resources or datasets based on a certain charactieristics. It it therefore important to be able to accurately represent or summarize such information, as well as track relations between entities. This section aims to illustrate how DATS model provides the relevant mechanisms to do so.

1.3.1 Tracking patient and specimen relationships

Relationships between materials matter. It is therefore important for the model to be able to represent information assessing sample / specimen origin and patient identity. For instance, in the context of longitudinal studies, repeated measure designs, where samples are collected or variables measured several times over the course of a study. The figure below shows the main properties of the DATS Material object, with associations to key biologically relevent entities such as:

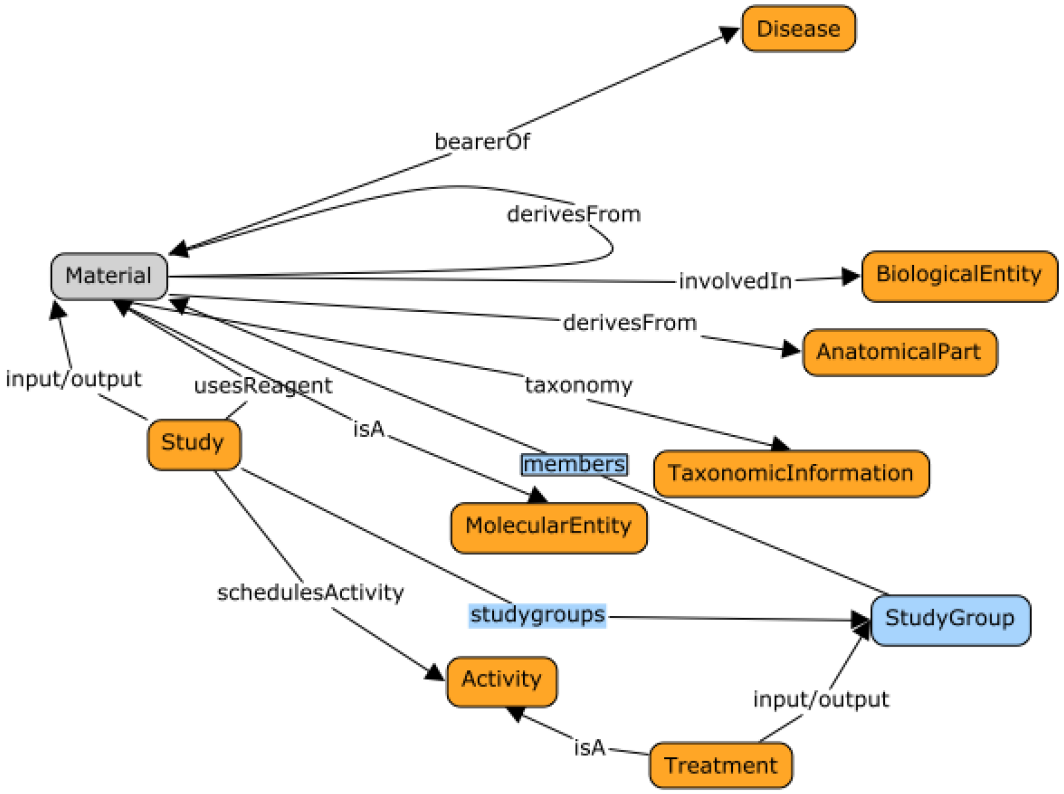
> Anatomical Part , Disease , Molecular Entity



Owing to awareness in resources such DO, GO, UBERON, the ease in integration and compatibility with biomedical ontologies should be highlighted.

1.3.2 Groups and sizes in the context of studies

For all datasets characterising "signal", the ability to identify, list and characterise study populations matters, as does the ability to capture descriptors for 'treatment' or 'perturbations'.

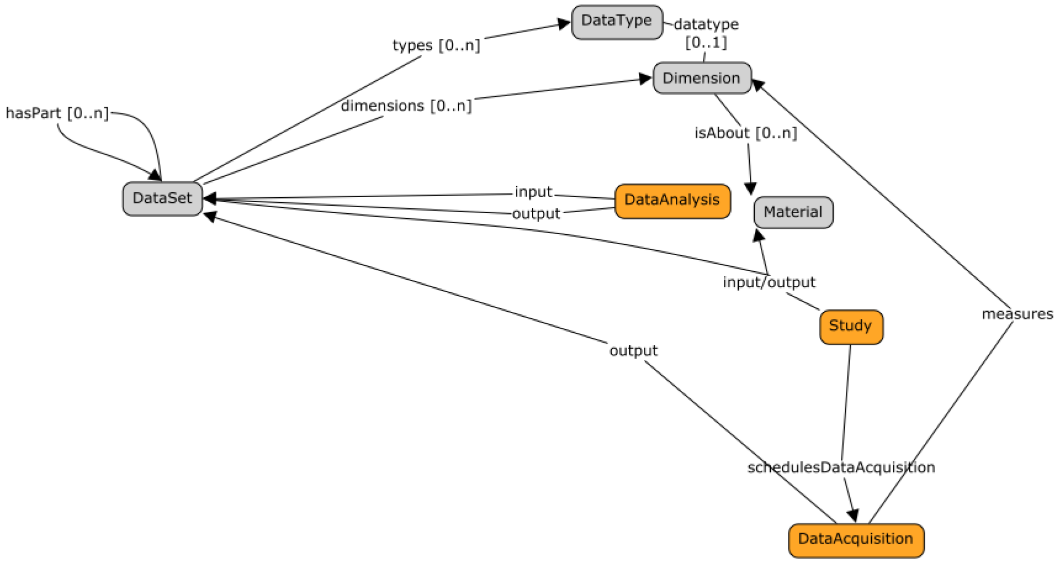


As shown in the figure above, the Data Study object allows the declaration and identification of groups (DATS Study Groups) of related materials as well as list all their members. The objects can be qualified with group size properties, allowing direct querying.

Note: While DATS model has been designed to enable granular representation, it does necessary follow that such granularity should always be used. Also, it is often the case, primary resources can not provide information to the extent required to perform the query case introduced at the top of the section.

1.4 DATS Measuring things:

This section describes the DATS objects for supporting the description of variables , dimensions and their relation to datasets.



The nature of the information available in a dataset can be recorded via the DATS Dimension entity. It is the object to use for reporting variables measured and for which data have been collected.

The DATS Dimension object can be qualitied using the DATS DataType entity.

The DATS *DataType* covers four aspects of a variable's nature: type of information (what the data is about), method (how the data was generated), platform (the instrumentation, software and reagents used to generate the data), and instrument (the specific device used to generate the data).

Importantly, it is key to remember that Dataset may be constitutive parts of another Dataset. Each of these dataset parts can be used to describe a particular aspect of a dataset in greater details. For instance, a dataset describing a multi-omics experiment may contain several datasets, one focusing on transcriptomics, one focusing on metabolomics and so on.

DATS.Dimension: meant to be used to report what data points are about in a dataset, their nature, their units.

DATS.Dimension should be typed (categorical, continuous)

DATS.Dimension used from the following DATS objects:

> DATS. Material .characteristics.Dimension
>
> DATS. DataAcquisition .measures.Dimension

1.5 Dataset Distribution

Where and How (can the dataset be accessed):

- Document DataSet Distribution options. This encompasses specifying:
  - data availability (boolean choice: available, unavailable)
  - data formats or mime-types ([terminology needs to be specified] 'resource: <https://github.com/lukaszsliwa/friendly_mime/blob/master/mimes.csv>'_)
  - data access conditions

- data compression (boolean choice: compressed, uncompressed)
- data encryption (boolean choice: encrypted, non-encrypted)
- data privacy protection (fully identifiable, pseudo-anonymized, full anonymized. . . .[terminology needs to be specified])

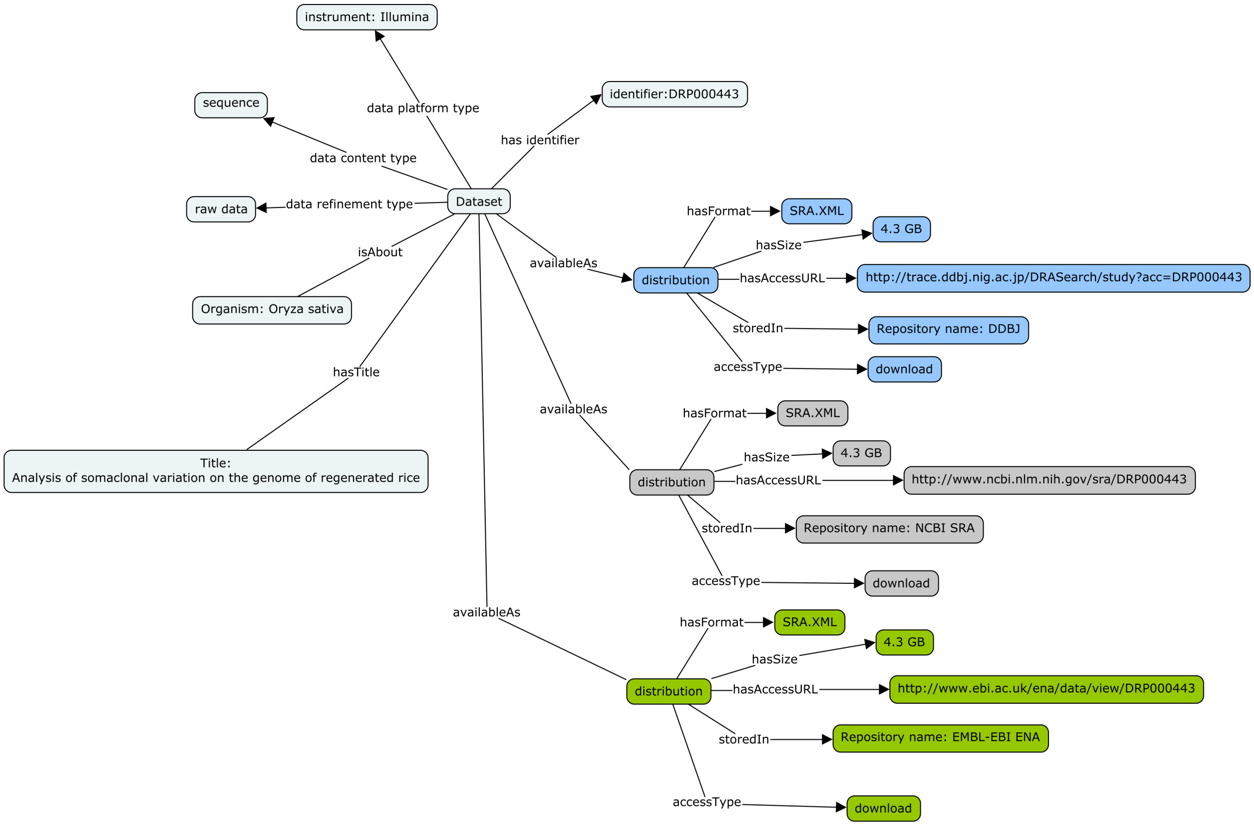
The image below provides an graphical overview of how to use Biocaddie DATS objects to encode information about dataset availability in a similar file format but from 3 distinct data repositories, each with it own access modalities.

The three INSDC sequence databases (DDBJ, SRA and ENA) exchange their data and provide the same datasets it in the three sites. Let's consider an example dataset.

The same Dataset identified by accession number DRP000443 can be accessed through the following 3 access URI pages:

- DDBJ:
- SRA:
- ENA:

While the distributions use the same Format, the accessURL are different as are the Repository but these distributions are all about the same dataset



The block below shows a snippet of a bioCADDIE DATS JSON document holding key information about dataset distribution. Note the link to *access information* and *data file format* information.

1.6 Dataset Creator(s)

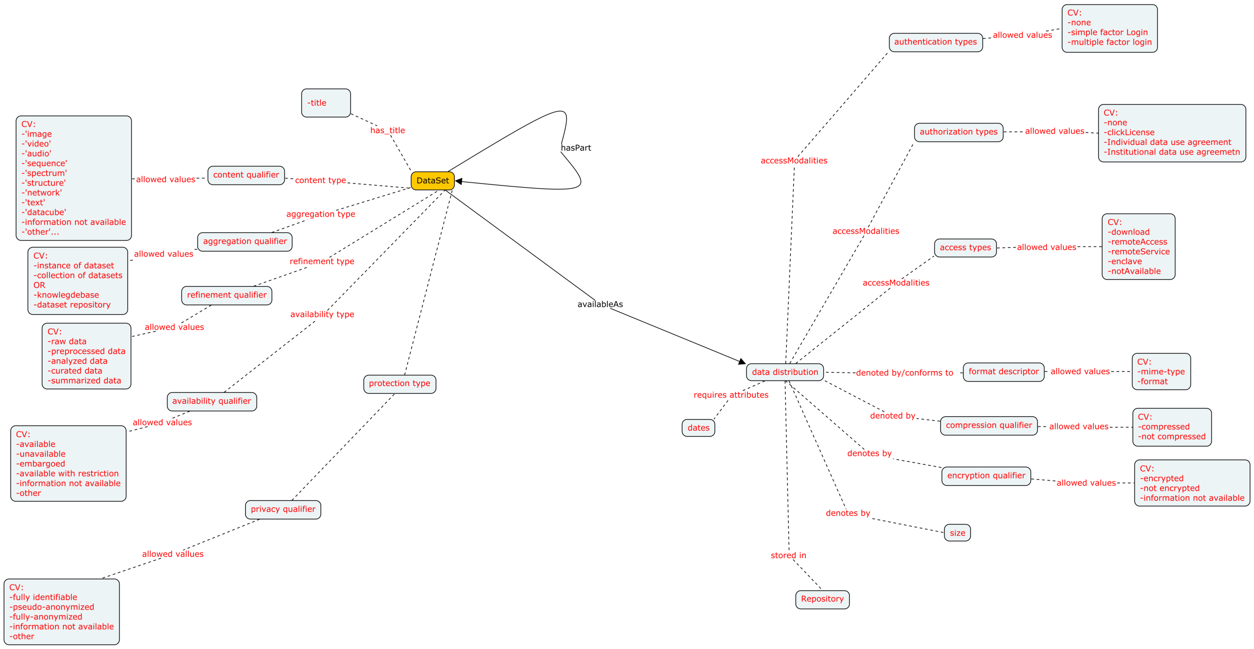
Who (produced the dataset):

- Document the Person(s) or Organization(s) which contributed to the creation of the Dataset.
- Document their roles (creator,curator,developer,funder,principal investigator. . . [terminology needs to be specified])

1.7 Dataset About

Describing what the dataset is about (i.e what was the scope, objective, materials) and providing information about the type of data associated with the given dataset:

- Document the nature of information available in a dataset through the Biocaddie **'data type'** object.



In this context, the *'data type'* required to annotate a DataSet should be viewed as a *content type* [terminology needs to be specified]). This encompasses the nature of the signal recorded in a dataset or information content of interest. For instance: gene expression data or phenotypic data, electronic health records But mime-type may be used. * chemical * sequence * spectrum * audio * image * video * . . .

but other descriptors may be used such as Biosharing, Scicrunch or re3data category/data domain descriptors.

- Data aggregation type:

  In the context of DataMed indexing, the information obtained from repositories may correspond to datasets served individually or may correspond to collections or records. As these 2 situations represent a very different metadata context, the Biocaddie DATS model allows to distinguish between the two cases.

- collection (as in 'collection of instances')
- singleton (as in 'individual instance')
- Data refinement type:

To describe the level of data processing associated with the data available from the dataset and its distributions. . . .[terminology needs to be specified])

- raw data
- preprocessed data

- analyzed data
- summarized data
- curated data
- reannotated data
- . . .
- data privacy protection type: (applicable only to human/clinical data)
  - fully identifiable none
  - pseudo-anonymized data
  - fully anonymized data
  - not information available
  - . . .
- Document the Material, object, scope and Biological Entities the dataset is about and their characteristics or properties.
- Document the nature of intervention and Treatment applied to the Material, if any or if applicable.
- Data Types and specific Platform

Currently, in DataMed, datasets can be search according to Data Type (.e.g Proteomics data) and/or by Platform (e.g. Illumina) DATS provides a mechanism via DataType object to qualify the nature of the data collected in a Dataset. The 4 facets/attributes allow to incrementally specify the type of information contained by the data and how it has been produced

- **data acquisition / method type:** This attribute allows to indicate the technique or technology , also known sometimes as data modality used to acquire the signal. For instance:
  - 'crystallography',
  - 'mass spectrometry'
  - 'nucleic acid sequencing',
  - 'computational simulation'
  - 'questionaire based survey'
  - 'nuclear magnetic resonance spectroscropy'
  - 'nuclear magnetic resonance imaging'
  - 'questionnaire'
  - . . .
- **platform/instrument type**
  - Agilent, Bruker,Affymetrix,Illumina,SeaHorse
  - HumanHap550v3.0
  - HumanExome-12 v1.1 BeadChip
  - Sentrix Human-6 Expression BeadChip
  - SureSelect Human All Exon v2 - 44Mb
  - HiSeq 2000
  - . . .

1.8 Dataset Provenance

In order to proceed with indexing a data source under bioCADDIE DataMed, it is essential to provide information about the actual source of information. This means unambiguously identifying the repository, the actual material from that resource used as input to the transformation allowing processing by DataMed software agents.

This falls under the provenance information section of the DATS for DataMed.

- identify the repository
- document the **url** or **filename** and address of the source information
- document the **date of last access** to the resource as input to the data transformation
- document the data transformation pipeline in the datamed infrastructure, ideally by pointed to the biocaddie github repository .

1.9 Frequently Asked Questions

1.9.1 Why are some properties (e.g. "title" and "description") included in both Dataset and DataDistribution?

When designing DATS we chose to be flexible and consider some redundancy by including properties in both Dataset as well as DatasetDistribution, even though in some cases it might be expected that a Dataset property should be inherited by their DatasetDistributions. We followed this approach to cover cases where repositories may have different information. For example, it would be possible that each DatasetDistribution has more information in its "description" on how the distribution was produced, adding more details to the general information in the corresponding Dataset.

CHAPTER 2

License:

BioCADDIE DATS is licensed under Creative Commons Attribution Share-Alike 4.0.

CHAPTER 3

Contributing:

If you wish to contribute to DATS and/or this documentation, please report issues in our tracker or contact us directly (agbeltran and proccaserra).

The different releases of DATS are available in the bioCADDIE Working Group 3 Github Repository, including documents and appendixes, JSON schemas, JSON-LD context files and JSON-LD instance files. Each release is preserved in the Zenodo repository and has its own persistent Digital Object Identifier (DOI). All releases in Zenodo can be accessed through the Zenodo DATS Community.

CHAPTER 4

Indices and tables:

- genindex
- modindex
- search