dats-doc Documentation

Release 0.1

Philippe Rocca-Serra and Alejandra Gonzalez-Beltran

Mar 27, 2017

Contents

CHAPTER 1

Introduction:

DATS, which stands for DAta Tag Suite, is a data description model designed and developed 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 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 documents accompanying each of the releases.

Table of Contents:

First Steps with DATS

This document gives an overview of the DATS components from a practical perspective, i.e. considering how DATS is used to describe a specific dataset considering a set of questions that determine the dataset provenance.

Who produced the dataset:

DATS records the *Person* (s) and *Organization* (s) associated with the dataset. In addition, it supports documenting their roles (e.g. creator, curator, developer, funder, principal investigator)

When was the data produced:

DATS records key *Date*(s) about the *Dataset*.

Each *Date* can specify its type, related to the event related to the key date (e.g. creation, update, validation, verification, deprecation of the dataset).

This mechanism of providing a generic *Date* indicating its type allows for extensions to new types of dates, which may be required in specific scenarios.

What is the dataset about:

DATS records the nature of information available in a dataset through the data type object.

Why was the data produced:

DATS supports to document the purpose, objective or hypothesis that gave origin to the dataset.

Where and *How* the dataset can be accessed:

DATS Model

Entity	Property	Definition
dataset	identifier	Primary iden
	relatedIdentifiers	Related iden
	alternateIdentifiers	Alternate id
	title	The name of
	types	A term, idea
	creators	The person(
	dates	Relevant dat
	distributions	The distribu
	dimensions	The differen
	isCitedBy	The relevant
	producedBy	A study proc
	isAbout	Different en
	hasPart	A Dataset th
	keywords	Tags associa
	acknowledges	The grant(s)
	extraProperties	Extra proper
DatasetDistribution		“A specific a
	identifiers	Primary iden
	alternateIdentifiers	Alternate id
	relatedIdentifiers	Related iden
	title	“The name o
	description	A textual na
	dates	“Relevant da
	“storedIn “	The data rep
	version	A release po
	accessModalities	The informa
	licenses	The terms o
	curationStatus	The level of
	conformsTo	A data stand
	format	The technica
	qualifiers	“One or mor
	“size “	The size of t
	unit	“The unit of

Entity	Property	Definition
	extraProperties	Extra proper
DataStandard		"A format, r
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	name	"The name o
	type	"The nature
	description	A textual na
	licenses	The terms o
	version	A release po
	extraProperties	Extra proper
DataRepository		A repository
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	name	The name of
	description	A textual na
	dates	Relevant dat
	scopes	"Information
	types	"A descripto
	licenses	The terms o
	version	"A release p
	publishers	The person(
	aggregatorOf	The DataRe
	accessModalities	The informa
	extraProperties	Extra proper
Software		"A digital en
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	name	The name of
	licenses	The terms o
	isUsedBy	The data acc
	manufacturer	The person o
	version	A release po
	extraProperties	Extra proper
Publication		A (digital) d
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	title	"The name o
	"dates "	"Relevant da
	type	"Publication
	publicationVenue	The name of
	authorsList	The list of a
	authors	The person(
	acknowledges	The grant(s)
	licenses	The terms o
	extraProperties	Extra proper

Entity	Property	Definition
IdentifiersInformation		Information
	identifier	A code uniq
	identifierSource	The identifie
AlternateIdentifiersInformation		Information
	alternateIdentifier	An identifier
	alternateIdentifierSource	The identifie
RelatedIdentifiersInformation		Information
	relatedIdentifier	An identifier
	relatedIdentifierSource	The identifie
	relationType	The type of
Annotation		“A pair of va
	“value “	A label or va
	ontologyTermIRI /suggested renaming = ValueIRI	The IRI of a
Date		“Information
	date	A date follo
Access		Information
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	landingPage	A web page
	accessURL	“A URL fro
	types	“Method to
	authorizations	Types of ver
	authentications	“Types of ve
	licenses	Terms of usa
	extraProperties	Extra proper
Grant		An allocated
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	name	The name of
	funds	The study or
	funders	The person(
	awardees	The person(
	extraProperties	Extra proper
License		“A legal doc
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	name	The name of
	version	The version
	creators	The person(
	extraProperties	Extra proper
Dimension		“A feature o
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	name	“The name o
	types	“A term, ide

Entity	Property	Definition
	partOf	The dataset(
	description	A textual na
	values	The actual c
	unit	“A reference
	“isAbout ‘‘	“A material
	extraProperties	Extra proper
	information	The measure
	method	The procedu
	platform	“The set of i
	instrument	The specific
	extraProperties	Extra proper
Material		“A physical
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	name	The name of
	derivesFrom	A material f
	bearerOfDisease	The patholo
	taxonomicInformation	The taxonon
	involvedInBiologicalEntity	A biological
	characteristics	The characte
	roles	The roles pl
	extraProperties	Extra proper
Person		A human be
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	fullName	“The first na
	firstName	The given na
	middleInitial	The first lett
	lastName	The person’s
	email	An electroni
	affiliations	The organiza
	roles	“The roles a
	extraProperties	Extra proper
	identifiers	Primary iden
	alternateIdentifiers	Alternate ide
	relatedIdentifiers	Related iden
	name	The name of
	abbreviation	“The shortna
	postalAddress	“The postal,
	roles	“The roles o
	extraProperties	Extra proper

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])

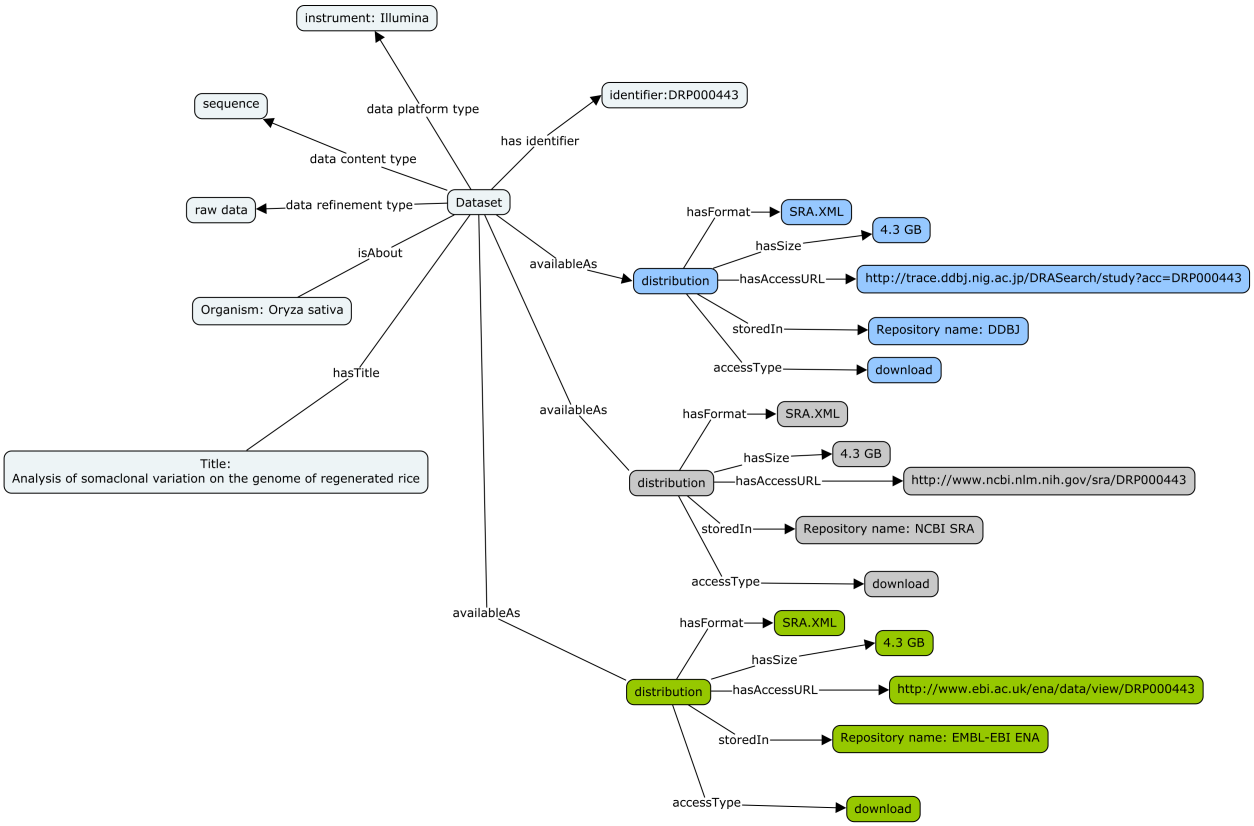
The image below provides an graphical overview of how to use Biocaddie DATS objects to encode information about dataset availability in a similar file format but from 3 distinct data repositories, each with it own access modalities.

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.

Dataset Creator(s)

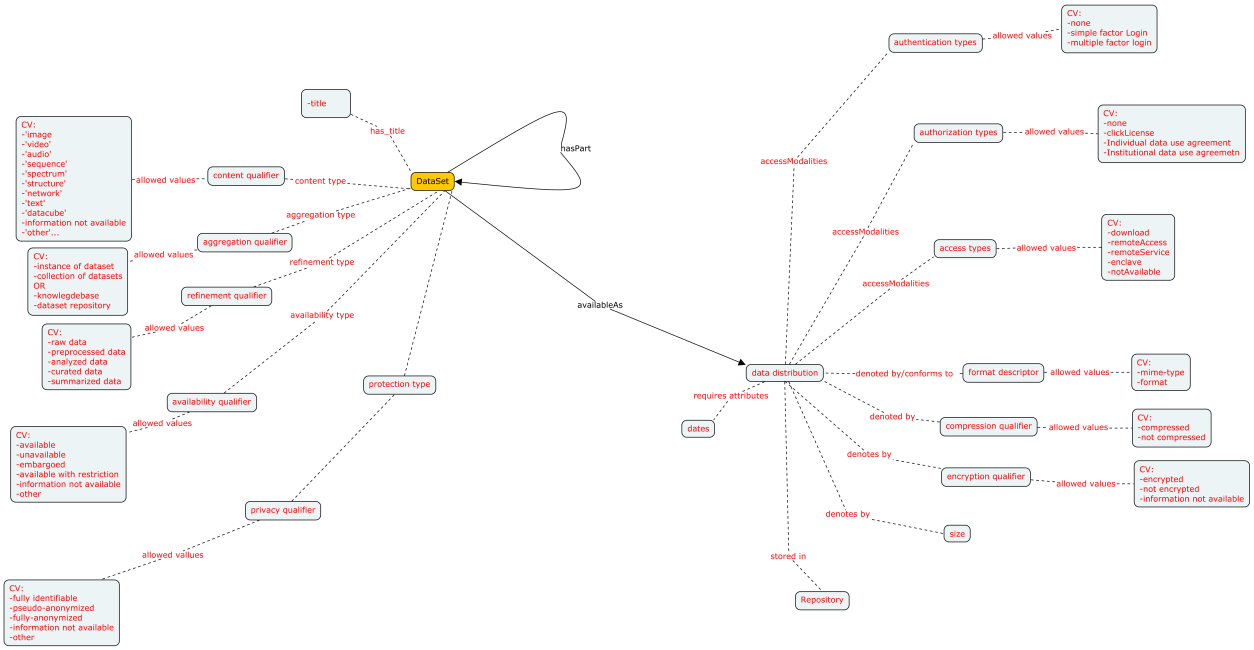
Who (produced the dataset):

- 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])

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
  - ...

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 .

Frequently Asked Questions

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.

CHAPTER 4

Indices and tables:

- genindex
- modindex
- search