DATA MARKET

AUSTRIA

tamarket.at

Data Market Austria Core Vocabulary

Deliverable number	*none*
Dissemination level	*Public*
Delivery date	*04.07.2018*
Status	*In development*
URL	*https://docs.google.com/document/d/1Xq7fjmYhzGNb0Qqc6DCcRNiK-D761dUnprW33wHiVlo/edit#*
Author(s)	*Martin Kaltenböck (SWC)* *Victor Mireles-Chavez (SWC)* *Hermann Fürntratt (JRS)* *Bernd Ivanschtitz (RSA)* *Bettina Rinnerbauer (DUK)* *Lörinc Thurnay (DUK)* *Thomas Lampoltshammer (DUK)* *Erwin Petz (ZAMG)* *Andreas Krimbacher (ZAMG)*

The Data Market Austria Project has received funding from the programme “ICT of the Future” of the Austrian Research Promotion Agency (FFG) and the Austrian Ministry for Transport, Innovation and Technology (Project 855404)

Abstract

The Data Maket Core Vocabulary (DMAV) provides classes and properties for describing datasets and services that are accessible on the Data Market Austria (DMA). It does not provide a formal, complete definition of all necessary dimensions for describing datasets, but rather sets out a consistent means by which dataset and service descriptions can be provided on the DMA. That way, it is possible for a potential user of DMA to find and retrieve datasets and services, as well as to judge their suitability for a particular purpose. This document describes the structure and usage of the Data Market Austria (DMA) Core Vocabulary for describing datasets (DMAV). It provides an overview on the main classes and properties that should be used within the IDS context in order to publish datasets and services. The namespace for DMAV is (URL not given at the moment - http://datamarket.at/2017/07/dmav/core#).

Nomenclature of the Metadata Core

Current Data set submission

```
dataset submission/
├── catalogue
│  ├── dataset 1
│  │  ├── data
│  │  │  └── example1.csv
│  │  ├── documentation
│  │  │  └── documentation1.odt
│  │  └── metadata
│  │      └── specific1.xml
│  └── dataset 2
│  │ ├── data
│  │ │  └── example2.csv
│  │ ├── documentation
│  │ │  └── documentation2.odt
│  │ └── metadata
│  │      └── specific2.xml
   └──metadata
    └── dcat.xml
```

Figure X: Dataset submission consisting of the Catalogue including the datasets and with associated metadata.

Data asset

Each dataset ingest process has a *process identifier* (UUID) and a corresponding working directory which is used as long as the ingest process is not finalized.

Data validation, transformation, and manipulation operations are executed in this working directory which contains the catalogue including the datasets and the metadata directory. When the ingest process is finished, the catalogue and metadata will be aggregated as a *data asset* and it must have an *internal identifier* (UUID) which is different from the process identifier.

Data assets have a life-cycle and can be changed and there is a distinction between two cases:

As illustrated in Figure X, an *edit* operation does not change the content in a way that it is necessary to assign a new persistent unique identifier. A *new version of the data asset* is created in this case. This can be a metadata correction or a format conversion which is supposed not to have any impact on the actual content. There will be a difference due to edit or conversion operations, however, the data asset can still be regarded as being “the same data asset” compared to the previous version.

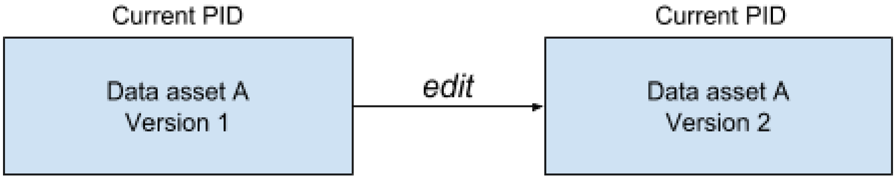


Fig X: An edit operation creates a new version of a data asset. The persistent unique identifier (PID) remains the same.

As illustrated in Figure X, a fork operation changes the content of the data asset and thereby creates a *new data asset derivative*. The difference to the source data asset(s) is different to a significant extent so that the new data asset cannot be regarded as being “the same data asset” compared to the previous version.

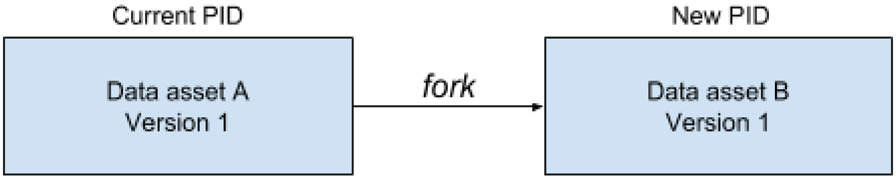


Fig X: A fork operation creates a new data asset which gets a new persistent unique identifier (PID)

The data asset derivative keeps the provenance relation to the original datasets it was derived from. Fork operations can also reference several source datasets.

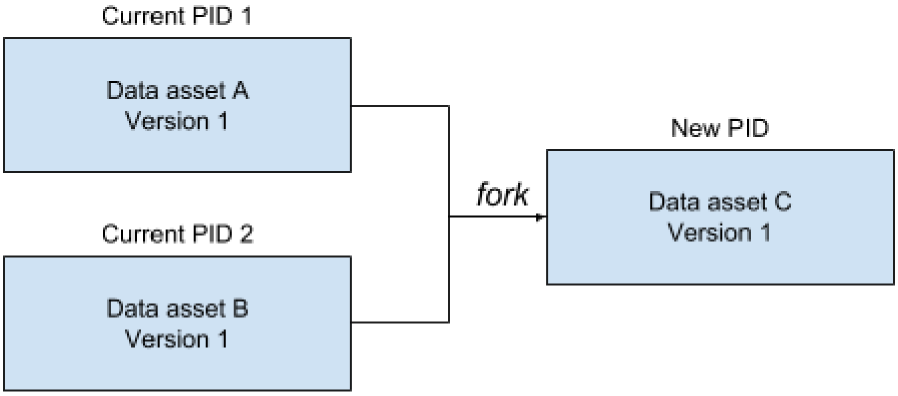


Fig X: A fork operation can reference multiple source data assets.

Namespaces

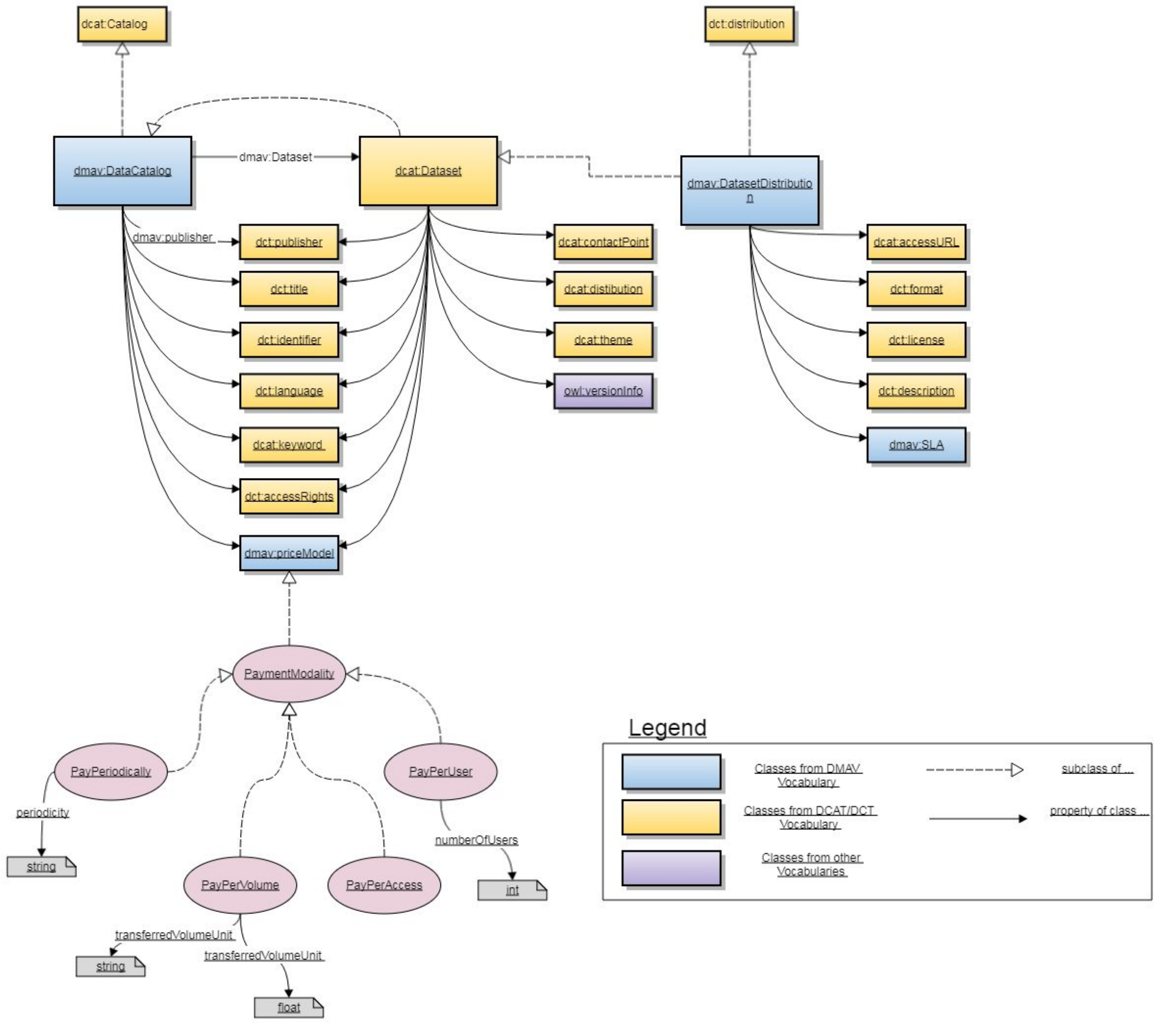
The DMAV uses a variation of the well known Data Catalog Vocabulary application profile (DCAT-AP). Besides the already mentioned DMAV namespace different other namespaces are use.

We reuse terms from various existing specifications. Classes and properties specified in the next sections have been taken from the following namespaces:

Prefix	Namespace
dcat	http://www.w3.org/ns/dcat#
dct	http://purl.org/dc/terms/
dctype	http://purl.org/dc/dcmitype/
foaf	http://xmlns.com/foaf/0.1/
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
skos	http://www.w3.org/2004/02/skos/core#
xsd	http://www.w3.org/2001/XMLSchema#
owl	http://www.w3.org/2002/07/owl#
schema	http://schema.org/
spdx	http://spdx.org/rdf/terms#
vcard	http://www.w3.org/2006/vcard/ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
dqv	http://www.w3.org/ns/dqv#
dmav	DMA Vocabularie (Pricing,...)

UML Class Diagram

A graphical overview on the vocabulary's structure is given in Figure 1. It illustrates the relevant classes and their alignment to existing vocabularies. Furthermore, it shows the properties that relate the classes to one another. To improve readability, we only depict the most relevant properties in the figure the DMA - Core.This section lists the classes, properties, and relationships contained in the application profile. The following sections indicate how these are represented in the RDF Schema.



Conformance

Since we use a the DCAT-AP we have to make sure that our defined metadata is compatible to the DCAT standard. Therefore, a data catalog conforms to DCAT if:

- It is organized into datasets and distributions.
- An RDF description of the catalog itself and its datasets and distributions is available (but the choice of RDF syntaxes, access protocols, and access policies is not mandated by this specification).
- The contents of all metadata fields that are held in the catalog, and that contain data about the catalog itself and its dataset and distributions, are included in this RDF description, expressed using the appropriate classes and properties from DCAT, except where no such class or property exists.
- All classes and properties defined in DCAT are used in a way consistent with the semantics declared in this specification.
- DCAT-compliant catalogs may include additional non-DCAT metadata fields and additional RDF data in the catalog's RDF description.

A **DCAT profile** is a specification for data catalogs that adds additional constraints to DCAT. A data catalog that conforms to the profile also conforms to DCAT. Additional constraints in a profile may include:

- A minimum set of required metadata fields
- Classes and properties for additional metadata fields not covered in DCAT
- Controlled vocabularies or URI sets as acceptable values for properties
- Requirements for specific access mechanisms (RDF syntaxes, protocols) to the catalog's RDF description

DMAV RDF Sample Structure

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:locn="http://www.w3.org/ns/locn#"
  xmlns:hydra="http://www.w3.org/ns/hydra/core#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcat="http://www.w3.org/ns/dcat#"
  xmlns:dct="http://purl.org/dc/terms/"
  xmlns:vcard="http://www.w3.org/2006/vcard/ns#"
  xmlns:dmav = “http://datamarket.at/2017/07/dmav/core#”
```

DMA Controlled Vocabulary

A controlled vocabulary is a restricted list of words or terms used for labeling, indexing or categorizing. It is controlled because only terms from the list may be used for the subject area covered by the controlled vocabulary. Most controlled vocabularies also have some form of cross-references pointing from one or more "non-preferred" terms to the designated "preferred" term.

Specific types of controlled vocabularies:

- Thesaurus: A thesaurus is a more structured kind of controlled vocabulary.
- Taxonomy: Has become a popular term now for any hierarchical classification or categorization system
- Ontology: Set of concepts with attributes and relationships between the various concepts that contain various meanings, all to define a domain of knowledge, and is expressed in a format that is machine-readable.

https://www.w3.org/standards/semanticweb/ontology
http://eurovoc.europa.eu/drupal/

RDF Class:	**dct:language**
Definition:	The language of the dataset.
Usage note:	● This overrides the value of the catalog language in case of conflict. ● If the dataset is available in multiple languages, use multiple values for this property. If each language is available separately, define an instance of dcat:Distribution for each language and describe the specific language of each distribution using dct:language (i.e. the dataset will have multiple dct:language values and each distribution will have one of these languages as value of its dct:language property).
Controlled Vocabulary:	dct:LinguisticSystem Resources defined by the Library of Congress (1, 2) should be used. If a ISO 639-1 (two-letter) code is defined for language, then its corresponding IRI should be used; if no ISO 639-1 code

	is defined, then IRI corresponding to the ISO 639-2 TODO: def if two-letters or three are used
RDF Example:	dct:language <http://id.loc.gov/vocabulary/iso639-1/en> ;

Beispiel(Metadata of Language):

URI	http://id.loc.gov/vocabulary/iso639-2/eng
iso639P1Code	en
iso639P3PCode	eng

RDF Class:	**dcat:keyword**
Definition:	A keyword or tag describing the dataset.
Usage note:	
Controlled Vocabulary:	rdfs:Literal We have to define a Keyword Catalog for our dmac vocabulary
RDF Example:	dcat:keyword "accountability","transparency" ,"payments" ;

RDF Class:	**dct:accessRights**
Definition:	A keyword or tag describing the dataset.
Usage note:	

Controlled Vocabulary:	dct:RightsStatement This property refers to information that indicates whether the Dataset is open data, has access restrictions or is not public. A controlled vocabulary with three members **(:public, :restricted, :non-public)** will be created and maintained by the Publications Office of the EU.
RDF Example:	`<dct:accessRights>public</dct:accessRights>`

RDF Class:	**dmav:pricing**
Definition:	Provides some information about the pricing system used by this catalogue; this price model is NOT valid for the single datasets included in the catalogue
Usage note:	
Controlled Vocabulary:	● :PayPerVolume ● :PayPerUser ● :PayPeriodically ● :PayOnce ● ...
RDF Example:	`<dmav:pricing>PayOnce</dmav:pricing>`

RDF Class:	**dcat:theme**
Definition:	The main category of the dataset & service. A dataset or service can have multiple themes.

Usage note:	
Controlled Vocabulary:	skos:Concept A new vocabulary for use in dcat:theme is being defined by the Publications Office of the European Union. The EU Data Theme Vocabulary will be available at the URI http://publications.europa.eu/resource/authority/data-theme, and described at the landing page http://publications.europa.eu/mdr/authority/data-theme.

RDF Example:	`<dcat:theme`
	`rdf:resource="http://eurovoc.europa.eu/100142"/>`
	` <dcat:theme>Earth Sciences</dcat:theme>`

RDF Class:	**dcat:themeTaxonomy**
Definition:	This property refers to a knowledge organization system used to classify the Catalogue's Datasets.
Usage note:	
Controlled Vocabulary:	skos:ConceptScheme The knowledge organization system (KOS) used to classify catalog's datasets.

RDF Example:	`<dcat:themeTaxonomy`
	`rdf:resource="http://example.org/skos/scheme/gemet" />`

RDF Class:	**dct:license**

Definition:	This property refers to the licence under which the Catalogue can be used or reused.
Usage note:	
Controlled Vocabulary:	dct:LicenseDocument A legal document describing the copyright license of the element. One way (recommended best practice) is to use **Creative Commons licenses** and to describe them in RDF with the Creative Commons Rights Expression Language (CC REL).
RDF Example:	`<dct:license rdf:resource="http://creativecommons.org/licenses/by/3.0/" />`

http://idi.fundacionctic.org/dcat-viewer/help

RDF Class:	**dct:accessRights**
Definition:	A keyword or tag describing the dataset.
Usage note:	
Controlled Vocabulary:	dct:RightsStatement This property refers to information that indicates whether the Dataset is open data, has access restrictions or is not public. A controlled vocabulary with three members **(:public, :restricted, :non-public)** will be created and maintained by the Publications Office of the EU.
RDF Example:	`<dct:accessRights>public</dct:accessRights>`

RDF Class:	**dmav:SLA**

Definition:	This property refers to the official commitment that prevails between a service provider and a client. Particular aspects of the service – quality, availability, responsibilities – are agreed between the service provider and the service user.
Usage note:	created new URI: dmav:SLA; should be a controlled vocabulary that needs to be specified. OPEN (to be discussed): is this a property of dataset OR distribution. THIS NEEDS to be specified in sub group involving also partners as Compass et al
Controlled Vocabulary:	● ● ...
RDF Example:	`<dmav:SLA>Text</dmav:SLA>`

RDF Class:	**dmav:priceModel**
Definition:	Provides some information about the price model used by this dataset
Usage note:	needs to be discussed: text field, versus controlled vocabulary - the model is unique, or? So cardinality = 1 (?); new URI created: dmav:PriceModel
Controlled Vocabulary:	● ● ...
RDF Example:	<dmav:**priceModel**>Text</dmav:**priceModel**>

RDF Class:	**dmav:SLA**

Definition:	This property refers to the official commitment that prevails between a service provider and a client. Particular aspects of the service – quality, availability, responsibilities – are agreed between the service provider and the service user.
Usage note:	created new URI: dmav:SLA; should be a controlled vocabulary that needs to be specified. OPEN (to be discussed): is this a property of dataset OR distribution. THIS NEEDS to be specified in sub group involving also partners as Compass et al
Controlled Vocabulary:	● ● ...
RDF Example:	`<dmav:SLA>Text</dmav:SLA>`

RDF Class:	**dmav:ServiceType**
Definition:	The type of service a customer can expect
Usage note:	
Controlled Vocabulary:	● Examples might be: 1. Batch Processing 2. Real time processing 3. Online Processing 4. Multiprocessing 5. Time sharing 6. Conversion converting data to another format. 7. Validation – Ensuring that supplied data is "clean, correct and useful." 8. Sorting – "arranging items in some sequence and/or in different sets." 9. Summarization – reducing detail data to its main points.

	10. Aggregation – combining multiple pieces of data. 11. Analysis – the "collection, organization, analysis, interpretation and presentation of data.". 12. Reporting – list detail or summary data or computed information. Query (Search) Calculation (Aggregation, Clustering, Classification) Content Delivery (Streaming) Decision Support (Recommender System) etc.
RDF Example:	`<dmav:ServiceType>Text</dmav:ServiceType>`

http://lov.okfn.org/dataset/lov/vocabs/dcat
https://www.w3.org/2013/dwbp/wiki/Making_controlled_vocabularies_accessible_as_URI_sets
http://id.loc.gov/vocabulary/iso639-2/eng.html