

The Data CaTalog metadata standard The ODI Athens Node

Bank of Greece Athens 08-11-2017

Attribution-ShareAlike 4.0 International









Session 1: Introduction to open data and metadata

Hopefully by the end of this session you will be able to

- Define Open Data
- Define metadata and dataset
- Explain how metadata standard can be used



OPEN DATA



Why now?

- policy mandates
- data and technical standar
- best practices and guidelines





what I need to know? ATHENS









The data spectrum





The data spectrum





The data spectrum











Δι@υγειαInstagram
Feed
Anonymized Health
RecordsAthens
Temperatures
Imetables
(KTEL)Imetables
Athens



Closed

Shared





CAN YOU GIVE AN OPEN DATA DEFINITION ?





Open ?

Open means **anyone** can **freely access**, **use**, **modify**, and **share** for **any purpose**

(subject, at most, to requirements that preserve provenance and openness)^{http://opendefinition.org}

ATHENS

Definition ?

Open data is information that is available for **anyone** to **use**, for **any purpose**, at **no cost**

- Open Data Institute



ATHENS

Definition

Open Data is Data that **anyone** can **access**, **use** and **share**

- Open Data Institute
- http://theodi.org/guide/whatopen-data Introduced in 2015





Is That All ?

There is something missing..

OPEN DATA

must be explicitly licensed





What is metadata, dataset and data distribution



What is metadata?

"Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information."

-- National Information Standards Organization http://www.niso.org/publications/press/UnderstandingMetadata .pdf

Metadata provides information enabling to make sense of **data** (e.g. documents, images, datasets), **concepts** (e.g. classification schemes) and **real-world entities** (e.g. people, organisations, places, paintings, products).



Do we really need a metadata standard?



When you manage a single or few datasets and you share them with colleagues you may not need a common metadata standard and you may use a common agreed schema

BUT when you want to share your heterogeneous datasets and contribute them to a data catalog then the dataset metadata (machine readable data description) becomes important







Data can be reused by applications: other can search and make use of your data

Datasets have to be found by applications
 Datasets have to be understood by applications

Datasets should be managed in data repositories/data catalogs

Data catalogs have to provide enough dataset metadata to applications to allow them to find and understand datasets





Types of metadata

Descriptive: describe a resource for discovery and identification

Structural: e.g. data models and reference data

Administrative: information to help resource management



Metadata schema



It is a standard set of data elements that is used for describing records. It establishes data elements and rules for their use.



What is important when creating a metadata schema for your data



Reuse properties from standard well accepted vocabularies. Some examples

- Dublin Core for published material (text, images), <u>http://</u> <u>dublincore.org/documents/dcmi-terms/</u>
- FOAF for people and organisations, <u>http://xmlns.com/foaf</u> /<u>spec/</u>
- SKOS for concept collections, <u>http://www.w3.org/TR/skos</u> <u>-reference</u>
- **ADMS** for interoperability assets, http://www.w3.org/TR/vocab-adms/
- Specific standard for datasets:
- Data Catalog Vocabulary DCAT, http://www.w3.org/TR/vocab-dcat/
- Specific usage of DCAT and other vocabularies to support interoperability of data portals across Europe.



How to provide metadata on the web?



Two approaches

```
<?xml version="1.0"?>
< |DOCTYPE user SYSTEM "users.dtd">
<user>
<student>
    <name>
      <firstname>Joe</firstname>
      <surname>Smith</surname>
      <title>Mr.</title>
      <username>smithj</username>
    </name>
    <contact>
      <address>
         <street>54 Maple Rise, Santry</street>
          <county>Dublin</county>
          <country>Ireland</country>
      </address>
      <email>smithj@dcu.ie</email>
                                                               e
    </contact>
                                                               e
    <programme active="true">
                                                               e
       <progname>M.Eng in Electronic Systems</progname>
      <code>MEN</code>
       <year>1</year>
    </programme>
    <module semester="2">
       <modid>EE557</modid>
    </module>
    <module semester="1">
       <modid>EE553</modid>
    </module>
 </student>
                     XML
```



ex:index.html	dc:creator	exstaff:85740 .
x:index.html	exterms:creation-date	"August 16, 1999" .
x:index.html	dc:language	"en" .

RDF (Triple-based approach)





What is a dataset?

According to W3C

Dataset is a collection of data, published or curated by a single agent, and available for access or download in one or more formats.

From: https://www.w3.org/TR/vocab-dcat/#class-distribution



What is a data distribution THENS

Represents 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: https://www.w3.org/TR/vocab-dcat/#class-distribution





Dataset description vocabularies ATHENS

• **DCAT** Vocabulary

- RDF vocabulary for describing any dataset
- Dataset can be stand alone or part of a catalog
- Metadata about dataset (collection) and related distributions

DataCube Vocabulary

- RDF Vocabulary for describing statistical datasets
- Useful for attaching metadata about the "data structure" of a dataset

• VOID vocabulary

- RDF vocabulary for expressing metadata about RDF datasets
- Useful especially for metadata related to RDF data services





Controlled vocabularies

What is a controlled vocabulary ATHENS

A controlled vocabulary is a predefined list of values to be used as values for a specific property in your metadata schema.

- In addition to careful design of schemas, the value spaces of metadata properties are important for the exchange of information, and thus interoperability.
- Common controlled vocabularies for value spaces make metadata understandable across systems.



How to use controlled vocabularies (1/3)



In element describing the topics: use a multilingual vocabulary with public URI for each term

Example: Use Eurovoc

<rdf:Description rdf:about="http://eurovoc.europa.eu/300">

<xl:altLabel rdf:resource="http://eurovoc.europa.eu/415040"/>
<s04:prefLabel xml:lang="da">international kredit</s04:prefLabel>
<s04:prefLabel xml:lang="sv">international kredit</s04:prefLabel>
<s04:prefLabel xml:lang="en">international credit</s04:prefLabel>
<s04:prefLabel xml:lang="de">international credit</s04:prefLabel>
<s04:prefLabel xml:lang="de">international kredit</s04:prefLabel>
</s04:prefLabel xml:lang="nl">international kredit</s04:prefLabel>
</s04:prefLabel xml:lang="nl">international kredit</s04:prefLabel></s04:prefLabel></s04:prefLabel xml:lang="nl">international kredit</s04:prefLabel></s04:prefLabel></s04:prefLabel xml:lang="nl">international kredit</s04:prefLabel></s04:prefLabel></s04:prefLabel xml:lang="nl">international kredit</s04:prefLabel></s04:prefLabel></s04:prefLabel xml:lang="nl">international kredit</s04:prefLabel></s04:prefLabel></s04:prefLabel xml:lang="nl">international kredit</s04:prefLabel></s04:prefLabel></s04:prefLabel xml:lang="nl">international kredit</s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:prefLabel></s04:pref



How to use controlled vocabularies (2/3)



In a Language element: use a standard code list

Example: Using ISO 639-1

<madsrdf:hasVariant> <madsrdf:Language> <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Variant"/> <madsrdf:variantLabel xml:lang="en">English</madsrdf:variantLabel> </madsrdf:Language> </madsrdf:hasVariant> <madsrdf:hasVariant>



How to use controlled vocabularies (3/3)



In an element for Country: use a standard ontology with public URIs for each value

Example: use Geonames ontology

dct:spatial <http://www.geonames.org/6695072> ;







LET'S TAKE A BREAK



Session 2: How to use DCAT

Hopefully by the end of this session you will be able to

- Explain the metadata lifecycle
- Name the main aspects of metadata quality
- **Define** DCAT and DCAT AP
- Explain how DCAT-AP can be used



CREATING, MAINTAINING, UPDATING AND PUBLISHING YOUR METADATA

Creating metadata



Metadata creation can be supported by (semi-)automatic processes.

- Document properties generated in (office) tools, e.g. creation date.
- Spatial and temporal information captured by cameras, sensors...
- Information from publication workflow, e.g. file location or URL

But other characteristics require human intervention:

- What is the resource about (e.g. linking to a subject vocabulary)?
- How can the resource be used (e.g. linking to a licence)?
- Where can I find more information about this resource (e.g. linking to a Web site or documentation that describes the





Maintaining metadata

Approaches for maintaining metadata need to be appropriate for the type of data that is being published.

Data does not change

- e.g. dataset with prices of goods for an older year
- metadata can be relatively stable.
 Changes (bulk conversions) can take place off-line when needed.

Data changes frequently

- e.g. sensor data
- metadata needs to be closely coupled to the data workflow and changes need to be practically instantaneous.



Updating metadata



Metadata operates in a global context that is subject to change!

- Organisation new departments are established, merge with others, responsibilities are handed over.
- Usage of the data new applications emerge around data.
- Reference data controlled vocabularies evolve and get linked.
- Data standards and technologies technology lifecycle is getting shorter all the time; what will tomorrow's Web look like?

Metadalsangedetense keptullontordetege, thenewtent, possible, takingointe...

account the available time and budget.



Storing metadata



Options for storing metadata

Depending on operational requirements, metadata can be embedded with the data or stored separately from the data.

- Embedding the metadata in the data (e.g. office documents, MP3, JPG, RDF data) embedding makes data exchange easier.
- Separating metadata from data (e.g. in a database), with links to corresponding data files makes management easier.

Depending on the availability of tools and requirements on performance and capacity, metadata can be stored in a 'classic' relational database or an RDF triple store.



Deleting metadata



Decommissioning or deletion of data happens, for example:

- When data is no longer necessary.
- When data is no longer valid.
- When data is wrong.
- When data is withdrawn by the owner/publisher

In that case the metadata should, contain information that the data was deleted, and if it was archived, how and where an archival copy can be requested.



Publishing metadata



'Open' publication: direct access on URIs

 This is the option most in line with the vision of Linked Open Data and allows the 'followyour-nose' principle.

Make your metadata available through a SPARQL endpoint

- This allows external systems to send queries to an RDF triple store.
- Requires knowledge about the schema used in the triple store.

Make your data available through a web service (REST based or SOAP)

- Can be used by software developers
- Requires good documentation including mappings to metadata standards
- Collaboration with metadata consumer

Deferred publication: access to exported file in RDF

- Produced by converting non-RDF data to RDF.
- Allows off-line bulk harvesting and caching of data collections.
- Allows implementation of access control.





Metadata quality



The quality and completeness of the description metadata of your datasets, directly affects their searchability and reuse.

From:

https://www.europeandataportal.eu/sites/default/files/d2.1.2_training_ module_1.4_introduction_to_metadata_management_en_edp.pdf



Metadata quality is about (1/ATHENS

- Accuracy: are the characteristics of the resource correctly reflected?
- Completeness: are all relevant characteristics of the resource captured (as far as practically and economically feasible and necessary for the application)?
- Availability: can the metadata be accessed now and over time into the future?





- Conformance: is the metadata conforming to a specific metadata standard or an Application Profile?
- Consistency: does the data not contain contradictions?
- Credibility and provenance: is the metadata based on trustworthy sources?



Metadata quality is about (3/ATHENS

- **Processability**: is the metadata properly machine-readable?
- Relevance: does the metadata contain the right amount of information for the task at hand?
- **Timeliness:** is the metadata corresponding to the actual (current) characteristics of the resource and is it published soon enough?



THE DATA CATALOG VOCABULARY (DCAT)



Some history



International Conference on Electronic Government EGOV 2010: Electronic Government pp 339-350 | Cite as

Enabling Interoperability of Government Data Catalogues

Authors

Authors and affiliations

Fadi Maali, Richard Cyganiak, Vassilios Peristeras

Conference paper

6	125	1.2k
Citations	Readers	Downloads

Part of the Lecture Notes in Computer Science book series (LNCS, volume 6228)

Abstract

Opening public sector information has recently become a trend in many countries around the world. Online government data catalogues with national, regional or local scope act as one-stop data portals providing descriptions of available government datasets. These catalogues though remain isolated. Potential benefits from federating geographically overlapping or thematically complementary catalogues are not realized. We propose an RDF Schema vocabulary as an interchange format among data catalogues and as a way of bringing them into the Web of Linked Data, where they can enjoy interoperability among themselves and with other deployed



ATHENS

The DCAT Model







DCAT Vocabulary 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#
vcard	http://www.w3.org/2006/vcard/ns#
xsd	http://www.w3.org/2001/XMLSchema#

DCAT makes extensive use of terms from other vocabularies



DCAT properties: catalog class





NODE

ATHENS



DCAT properties: Catalog record class

A record in a data catalog, describing a single dataset. (Optional)

Property	Usage
dct: title	Name of the given record
dct: description	Free text description of the record
dct: issued	The date of listing the corresponding dataset in the catalog.
dct: modified	Most recent date on which the catalog entry was changed, updated or modified.
dct: primarytopic	Links the catalog record to the dcat:Dataset resource described in the record.



DCAT properties: dataset class



Actual dataset as published by the dataset publisher

Property	Usage
dct: title	A name of the dataset
dct: description	Free text description of the dataset
dct: issued	The date of formal publication of the dataset
dct: modified	Most recent date on which the dataset was changed, updated or modified.
dct: language	The language of the dataset
dct: publisher	An entity responsible for making the dataset available.
dct: acrualPeriodicity	The frequency at which dataset is published.
dct: identifier	A unique identifier of the dataset
dct: spatial	Spatial coverage of the dataset
dct: temporal	Temporal period that the dataset covers
dcat: theme	The main category of the dataset. A dataset can have multiple themes.
dcat: keyword	A keyword or tag describing the dataset.
dcat: contactPoint	Link a dataset to relevant contact information which is provided using VCard
deat distribution	Connect a dataset to its distributions



DCAT properties: distribution class

Represents a specific available form of a dataset.

Property	Usage
dct: title	Name of the distribution
dct: description	Free text description of the distribution
dct: issued	Date of formal publication
dct: modified	Most recent date on which the distribution was changed, updated or modified.
dct: license	This links to the license document under which the distribution is made available.
dct: rights	Information about rights held in and over the distribution.
dcat: accessURL	A landing page, feed, SPARQL endpoint or other type of resource that gives access to the distribution of the dataset
dcat: downloadURL	A file that contains the distribution of the dataset in a given format
dcat: byteSize	The size of a distribution in bytes.
dcat: mediaType	The media type of the distribution
dct: format	format of the distribution



DCAT Metadata Application Profile



A **DCAT application 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

--- https://www.w3.org/TR/vocab-dcat/





Supported by a community

ISA² Programme

https://joinup.ec.europa.eu/solution/dcat-application-profile-dataportals-europe







DCAT – AP extensions

Examples of National extensions

- DCAT-AP Sweden
- DCAT-AP Belgium
- DCAT-AP Germany
- DCAT-AP Spain

Domain extensions

- **GEO/DCAT-AP** in collaboration with JRC team responsible for INSPIRE Directive
- **STAT/DCAT-AP** in collaboration with EUROSTAT
- recently some new extensions have been developed for transportation, agrifood, chatbots, scientific data





Exchanging your data descriptions (metadata)

Mapping metadata to a common metadata vocabulary



- Mapping of properties: e.g. Dataset name [] DC Title, description
- Mapping of controlled vocabularies: e.g. Language, topics, format



DCAT-AP as the common metadata schema to exchange dataset information ATHENS

The DCAT-AP can be used to collect (harvest) the metadata from all data portals enabling the **Open Data** Interoperabilit V







LET'S TRY A MAPPING



Bank of Greece





http://www.bankofgreece.gr/BoGDocurnents/Νέοι Πίνακες Τιμών Κατοικιών full.pdf







Define metadata standard, dataset and DCAT

<u>Name</u> the most important aspects of metadata quality

Describe how DCAT can be used to describe and exchange metadata for a dataset

References



- W3C, The DCAT Specification, <u>https://www.w3.org/TR/vocab-dcat/</u>
- Makx Dekkers, Michiel De Keyzer, Nikolaos Loutas and Stijn Goedertier, Introduction to metadata management
- Valeria Pesce, Dataset description: DCAT and other vocabularies, https://www.slideshare.net/valeriap/datasetdescription-dcat-and-other-vocabularies





LET'S TAKE A BREAK