



# OPEN DATA SUPPORT

## Training Module 1.4

# Introduction to metadata management



*PwC firms help organisations and individuals create the value they're looking for. We're a network of firms in 158 countries with close to 180,000 people who are committed to delivering quality in assurance, tax and advisory services. Tell us what matters to you and find out more by visiting us at [www.pwc.com](http://www.pwc.com). PwC refers to the PwC network and/or one or more of its member firms, each of which is a separate legal entity. Please see [www.pwc.com/structure](http://www.pwc.com/structure) for further details.*

# Presentation metadata

Open Data Support is funded by the European Commission under SMART 2012/0107 'Lot 2: Provision of services for the Publication, Access and Reuse of Open Public Data across the European Union, through existing open data portals'(Contract No. 30-CE-0530965/00-17).

© 2014 European Commission

---

***This presentation has been created by PwC***

## ***Authors:***

**Makx Dekkers, Michiel De Keyzer, Nikolaos Loutas  
and Stijn Goedertier**

## ***Disclaimers***

1. The views expressed in this presentation are purely those of the authors and may not, in any circumstances, be interpreted as stating an official position of the European Commission. The European Commission does not guarantee the accuracy of the information included in this presentation, nor does it accept any responsibility for any use thereof. Reference herein to any specific products, specifications, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favouring by the European Commission. All care has been taken by the author to ensure that s/he has obtained, where necessary, permission to use any parts of manuscripts including illustrations, maps, and graphs, on which intellectual property rights already exist from the titular holder(s) of such rights or from her/his or their legal representative.
2. This presentation has been carefully compiled by PwC, but no representation is made or warranty given (either express or implied) as to the completeness or accuracy of the information it contains. PwC is not liable for the information in this presentation or any decision or consequence based on the use of it.. PwC will not be liable for any damages arising from the use of the information contained in this presentation. The information contained in this presentation is of a general nature and is solely for guidance on matters of general interest. This presentation is not a substitute for professional advice on any particular matter. No reader should act on the basis of any matter contained in this publication without considering appropriate professional advice.

---

## *Learning objectives*

By the end of this training module you should have an understanding of:

- What metadata is;
- The terminology and objectives of metadata management;
- The use of controlled vocabularies for metadata;
- The creation and publication of description metadata of datasets on the EU ODP.

---

# *Content*

This module contains ...

- An explanation of what is metadata;
- An outline of how to create and publish metadata on the EU ODP.

*Find more on: [training.opendatasupport.eu](http://training.opendatasupport.eu)*

---

# What is metadata?

*Definition, examples and reusable standards.*

---

# *What is (description) 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>

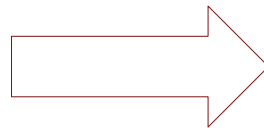
Metadata provides information enabling to make sense of **data** (e.g. documents, images, datasets), **concepts** (e.g. classification schemes) and **real-world entities** (e.g. people, organisations, places, paintings, products).

# Examples of metadata

Label



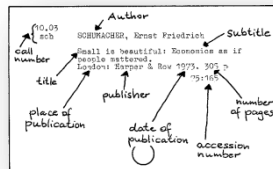
Provides metadata on



Can



Catalogue card



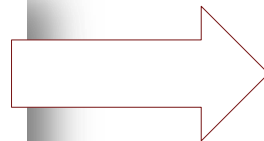
Book



Dataset description (DCAT)

```
weather1-7 a dcat:Dataset ;
dct:title "Measurements from weather stations 1-7" ;
dct:description "Data from seven weather stations
showing temperature, humidity,
wind direction and wind speed" ;
dct:modified "2013-07-01" ;
dct:publisher <http://myweather.com/id/myweather> ;
dcat:keyword "weather" ;
dcat:landingpage <http://myweather.com/stations1-7.html> ;
dcat:distribution :weatherdata-xlsx
.

weatherdata1-7-xlsx a dcat:Distribution ;
dct:format <http://publications.europa.eu/resource/authority/file-type/XLSX> ;
dct:licence <http://creativecommons.org/licenses/by/4.0/> ;
dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
.
```



Dataset

	Temp. °C	Humidity %	Wind direction	Wind speed km/h
Station 1	18.1	60	WSW	18
Station 2	17.5	59	WSW	20
Station 3	18.2	55	SW	22
Station 4	19.0	62	SW	18
Station 5	18.0	65	WSW	19
Station 6	18.2	63	SSW	21
Station 7	17.9	61	SW	22

# *Example: description of an open dataset with the DCAT-AP*

Description of the  
Catalogue

```
:catalog
  a dcat:Catalog ;
  dct:title "Imaginary Catalog" ;
  rdfs:label "Imaginary Catalog" ;
  foaf:homepage <http://example.org/catalog> ;
  dct:publisher :transparency-office ;
  dct:language <http://id.loc.gov/vocabulary/iso639-1/en> ;
  dcat:dataset :dataset-001 , :dataset-002 , :dataset-003 ;
  .
```

Description of the  
Dataset

```
:dataset-001
  a dcat:Dataset ;
  dct:title "Imaginary dataset" ;
  dcat:keyword "accountability","transparency" ,"payments" ;
  dct:issued "2011-12-05"^^xsd:date ;
  dct:modified "2011-12-05"^^xsd:date ;
  dct:publisher :finance-ministry ;
  dct:language <http://id.loc.gov/vocabulary/iso639-1/en> ;
  dcat:distribution :dataset-001-csv ;
  .
```

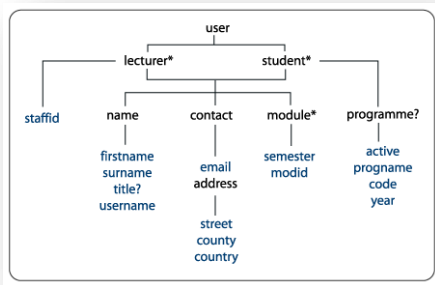
Description of the  
Distribution

```
:dataset-001-csv
  a dcat:Distribution ;
  dcat:downloadURL <http://www.example.org/files/001.csv> ;
  dct:title "CSV distribution of imaginary dataset 001" ;
  dcat:mediaType "text/csv" ;
  dcat:byteSize "5120"^^xsd:decimal ;
  .
```



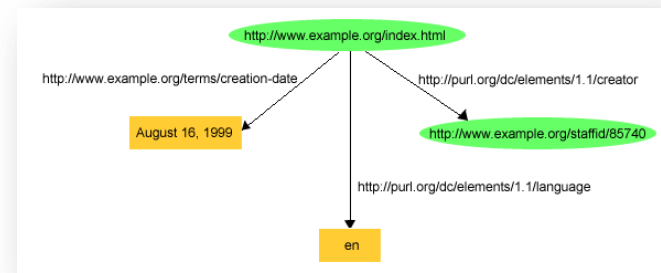
# Two approaches for providing metadata on the Web

## XML (Tree/container approach)



```
<?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>
        <country>Dublin</country>
        <country>Ireland</country>
      </address>
      <email>smithj@dcu.ie</email>
    </contact>
    <programme active="true">
      <programe>M.Eng in Electronic Systems</programe>
      <code>MEN</code>
      <year>1</year>
    </programme>
    <module semester="2">
      <modid>EE557</modid>
    </module>
    <module semester="1">
      <modid>EE553</modid>
    </module>
  </student>
```

## RDF (Triple-based approach)



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

---

## *Reuse existing vocabularies for providing metadata to your datasets*

*DCAT application profile for data portals in Europe,*  
[http://joinup.ec.europa.eu/asset/dcat\\_application\\_profile/description](http://joinup.ec.europa.eu/asset/dcat_application_profile/description)

- Based on DCAT – a W3C Recommendation  
<http://www.w3.org/TR/vocab-dcat/>
- Defines **mandatory**, **recommended** and **optional** classes and properties
- Recommends a number of **controlled vocabularies** for assigning values to properties, e.g. Eurovoc for dcat:theme.
- Currently implemented in the context of Open Data Support;
- A number of Member States are considering its adoption;
- The metadata model of the EU ODP will also converge.

---

# Controlled vocabularies

*Using thesauri, taxonomies and standardised lists of terms for assigning values to metadata properties.*

---

# *What are controlled vocabularies?*

*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.

# Which controlled vocabulary to be used for which type of property

- Use **code lists** as controlled vocabulary for free text or “string” properties.
- Example DCAT-AP property:

keyword/ tag	dcat:keyword	rdfs:Literal	This property contains a keyword or tag describing the dataset.
--------------	--------------	--------------	---

- Example code list - ObjectInCrimeClass (ListPoint)

Code	Meaning
1	<a href="#">Used</a>
2	<a href="#">Taken Without Consent</a>
3	<a href="#">Eliminated</a>
4	<a href="#">Not Involved</a>
5	<a href="#">Involved</a>
6	<a href="#">Crime Weapon</a>
7	<a href="#">Handled / Received</a>
11	<a href="#">Exhibit</a>

- Use **concepts identified by a URI** for reference to “things”.
- Example DCAT-AP property:

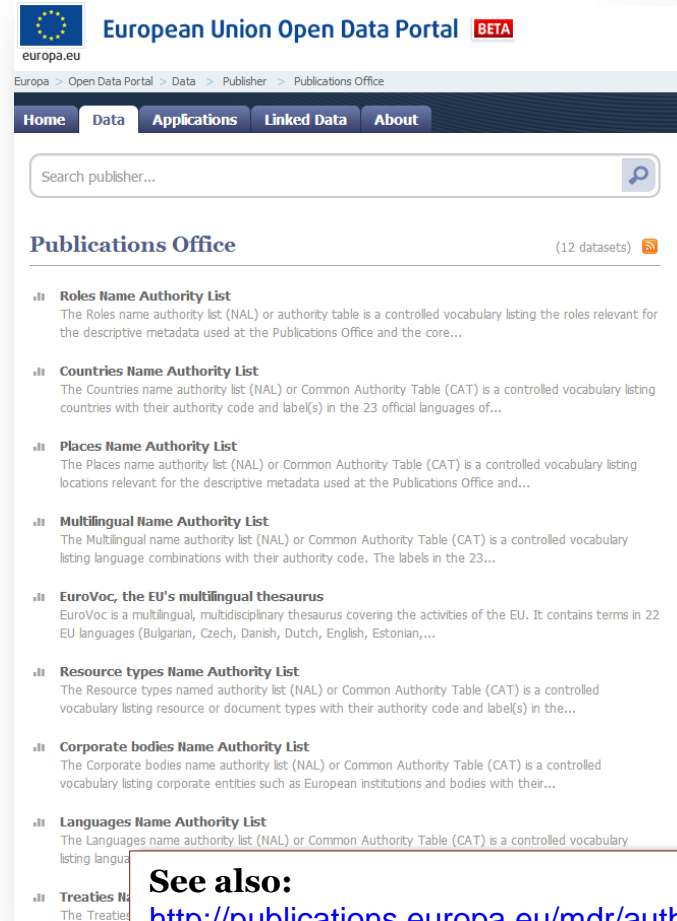
theme/ category	dcat:theme, subproperty of dct:subject	skos:Concept	This property refers to a category of the dataset.
-----------------	--	--------------	--

- Example taxonomy with terms having a URI - 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">internationell kredit</s04:prefLabel>
  <s04:prefLabel xml:lang="en">international credit</s04:prefLabel>
  <s04:prefLabel xml:lang="de">internationaler Kredit</s04:prefLabel>
  <s04:prefLabel xml:lang="nl">internationaal krediet</s04:prefLabel>
</rdf:Description>
```

# Example – Publications Office's Named Authority Lists

- The Named Authority Lists offer reusable controlled vocabularies for:
  - Countries
  - Corporate bodies
  - File types
  - Interinstitutional procedures
  - Languages
  - Multilingual
  - Resource types
  - Roles
  - Treaties



The screenshot shows the European Union Open Data Portal (BETA) interface. The page title is "Publications Office" with a sub-header "(12 datasets)". Below the title is a list of Named Authority Lists (NALs) with brief descriptions:

- Roles Name Authority List**: The Roles name authority list (NAL) or authority table is a controlled vocabulary listing the roles relevant for the descriptive metadata used at the Publications Office and the core...
- Countries Name Authority List**: The Countries name authority list (NAL) or Common Authority Table (CAT) is a controlled vocabulary listing countries with their authority code and label(s) in the 23 official languages of...
- Places Name Authority List**: The Places name authority list (NAL) or Common Authority Table (CAT) is a controlled vocabulary listing locations relevant for the descriptive metadata used at the Publications Office and...
- Multilingual Name Authority List**: The Multilingual name authority list (NAL) or Common Authority Table (CAT) is a controlled vocabulary listing language combinations with their authority code. The labels in the 23...
- EuroVoc, the EU's multilingual thesaurus**: EuroVoc is a multilingual, multidisciplinary thesaurus covering the activities of the EU. It contains terms in 22 EU languages (Bulgarian, Czech, Danish, Dutch, English, Estonian,...
- Resource types Name Authority List**: The Resource types named authority list (NAL) or Common Authority Table (CAT) is a controlled vocabulary listing resource or document types with their authority code and label(s) in the...
- Corporate bodies Name Authority List**: The Corporate bodies name authority list (NAL) or Common Authority Table (CAT) is a controlled vocabulary listing corporate entities such as European institutions and bodies with their...
- Languages Name Authority List**: The Languages name authority list (NAL) or Common Authority Table (CAT) is a controlled vocabulary listing languages...
- Treaties Name Authority List**: The Treaties name authority list (NAL) or Common Authority Table (CAT) is a controlled vocabulary listing treaties...

See also:

<http://publications.europa.eu/mdr/authority/>

# *EuroVoc for labelling the themes of datasets*

- Managed by the Publications Office
- Thesaurus covering the activities of the EU
- Terms in 23 EU languages
- Users include
  - the European Parliament
  - the Publications Office
  - national and regional parliaments and governments in Europe
  - private users around the world



The screenshot shows the EuroVoc website interface. At the top, there is a logo for EuroVoc, the Multilingual Thesaurus of the European Union. Below the logo, a navigation bar indicates the current location: Europa > EuroVoc homepage > Domains and MT. The main content area is divided into several sections:

- Content language:** A dropdown menu is set to '(en) English'.
- Simple search:** A search input field with a search button.
- Advanced search:** A link to access more search options.
- Browse:** A link to browse the subject-oriented version.
- Download:** A list of download options: By domain, Permuted alphabetical, Multilingual list, Alphabetical index, and SKOS/XML.
- Your proposals:** A link to contribute or view new approved concepts.

On the right side, there is a section titled **Domains** with a list of categories and their sub-themes:

- 04 POLITICS
- 08 INTERNATIONAL RELATIONS
  - 0806 international affairs
  - 0811 cooperation policy
  - 0816 international balance
  - 0821 defence
- 10 EUROPEAN UNION
- 12 LAW
- 16 ECONOMICS
  - 1606 economic policy
  - 1611 economic growth
  - 1616 regions and regional policy
  - 1621 economic structure
  - 1626 national accounts
  - 1631 economic analysis
- 20 TRADE
- 24 FINANCE

**See also:**

<http://eurovoc.europa.eu/>

---

Creating and publishing  
description metadata of  
datasets on the EU ODP



---

# *Metadata management is important*

## *Metadata needs to be managed to ensure ...*

- **Availability:** metadata needs to be stored where it can be accessed and indexed so it can be found.
- **Quality:** metadata needs to be of consistent quality so users know that it can be trusted.
- **Persistence:** metadata needs to be kept over time.
- **Open License:** metadata should be available under a public domain license to enable its reuse.

## *The metadata lifecycle is **larger** than the data lifecycle:*

- Metadata may be **created before data is created** or captured, e.g. to inform about data that will be available in the future.
- Metadata needs to be **kept after data has been removed**, e.g. to inform about data that has been decommissioned or withdrawn.

---

# ***Creating and publishing your metadata on the EU ODP***

## *Manually creating your metadata using a spreadsheet template*

- **Use a spreadsheet template** that conforms to the metadata model of the EU ODP in order to create description metadata for your datasets.

## *Metadata creation using (semi-)automatic processes*

- **Develop an exporter** that exports the description metadata of your datasets from your database/system in a format that conforms to the requirements of the EU ODP.
- **Develop a screen-scraping/harvester** that collects the description metadata of your datasets from your portal and transforms it in a format that conforms to the requirements of the EU ODP.

---

## *Updating your metadata – planning for change*

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

- **Organisation** – 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?

*The description metadata of your datasets on the EU ODP needs to be kept up-to-date to the extent possible, taking into account the available time and budget.*

---

## *Storing your metadata*

- The description metadata of your datasets to be published on the EU ODP should be stored separate from the data – but should be linked to it.
- This makes metadata management –including sharing – easier.
- Depending on the availability of tools and requirements on performance and capacity, metadata can be stored in a **‘classic’ relational database**, a file on a Web location or an **RDF triple store**.

---

## *Conclusions*

- Description metadata provides information on your datasets.
- The quality of the description metadata directly affects the discoverability and reuse of your datasets.
- A structured approach should be followed for metadata management.
- The metadata lifecycle extends the lifecycle of datasets (metadata before publication and after deletion).
- Homogenised metadata enable the operation of metadata brokers, which can in turn lower the access barriers to your resources, leading to improved visibility and discoverability, and thus increasing their reuse potential.

---

## *Group exercise and questions*



<http://www.visualpharm.com>

In groups of two, select one dataset from your institution and describe it with the DCAT Application Profile.



<http://www.visualpharm.com>

Does your organisation maintain a minimum set of metadata to be provided together with its datasets?



<http://www.visualpharm.com>

Do you have any data and/or metadata governance methodology at the corporate level?

---

Thank you!

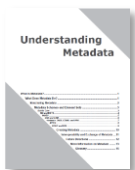
...and now YOUR questions?

# References

- NISO. Understanding Metadata. <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>
- W3C. RDF Primer. <http://www.w3.org/TR/rdf-primer/>
- [http://gondolin.rutgers.edu/MIC/text/how/catalog\\_glossary.htm](http://gondolin.rutgers.edu/MIC/text/how/catalog_glossary.htm)
- Dublin Core. Example XML Schema. <http://dublincore.org/schemas/xmls/qdc/dc.xsd>
- Dublin Core, Example RDF Schema. <http://dublincore.org/2012/06/14/dcterms.rdf>
- The ISA Programme. DCAT Application Profile for Data Portals in Europe - Final Draft. [https://joinup.ec.europa.eu/asset/dcat\\_application\\_profile/asset\\_release/dcat-application-profile-data-portals-europe-final-draft](https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final-draft)
- European Data Portal. <http://open-data.europa.eu/en/data/dataset?q=Name+Authority+List&op=>
- Publications Office. Countries Name Authority List. <http://open-data.europa.eu/en/data/dataset/2nM4aG8LdHG6RBMumfkNzQ>



## *Further reading*



Understanding Metadata, NISO.

<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>



Ben Jareo and Malcolm Saldanha. The value proposition of a metadata driven data governance program. Best Practices Metadata. May 2012.

[https://community.informatica.com/mpresources/Communities/IW2012/Docs/bos\\_30.pdf](https://community.informatica.com/mpresources/Communities/IW2012/Docs/bos_30.pdf)



John R. Friedrich, II. Metadata Management Best Practices and Lessons Learned. The 10th Annual Wilshire Meta-Data Conference and the 18th Annual DAMA International Symposium. April 2006.

<http://www.metaintegration.net/Publications/2006-Wilshire-DAMA-MetaIntegrationBestPractices.pdf>

## *Related initiatives*



Metadata Management. Trainer screencasts, <http://managemetadata.com/screencasts/msa/>



MIT Libraries. Data Management and Publishing. Reasons to Manage and Publish Your Data, <http://libraries.mit.edu/guides/subjects/data-management/why.html>



ISA Programme. DCAT Application Profile for European Data Portals, [https://joinup.ec.europa.eu/asset/dcat\\_application\\_profile/description](https://joinup.ec.europa.eu/asset/dcat_application_profile/description)



Generating ADMS-based descriptions of assets using Open Refine RDF, <https://joinup.ec.europa.eu/asset/adms/document/generate-adms-asset-descriptions-spreadsheet-refine-rdf>



The Dublin Core Metadata Initiative, <http://dublincore.org/>

# *Be part of our team...*

## *Find us on*



[Open Data Support](http://www.slideshare.net/OpenDataSupport)

<http://www.slideshare.net/OpenDataSupport>



[Open Data Support](http://goo.gl/y9ZZI)

<http://goo.gl/y9ZZI>

## *Follow us*



[@OpenDataSupport](https://twitter.com/OpenDataSupport)

## *Join us on*



joinup

<http://www.opendatasupport.eu>

## *Contact us*

[contact@opendatasupport.eu](mailto:contact@opendatasupport.eu)