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.

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 for further details.
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).

© 2013 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 different dimensions of metadata quality;
• The use of controlled vocabularies for metadata;
• Metadata exchange and aggregation;
• Metadata management in Open Data Support.
Content

This module contains...

• An explanation of what is metadata;
• An outline of the metadata lifecycle;
• An introduction to metadata quality;
• An overview of the metadata management and exchange approach implemented by Open Data Support through the Open Data Interoperability Platform.
What is metadata?

Definition, examples and reusable standards.
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


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
Catalogue card

Dataset description (DCAT)

Can
Book

Dataset
Two approaches for providing metadata on the Web

**XML (Tree/container approach)**

```xml
<user>
  <name><firstname>Jim</firstname> <surname>Smith</surname> <title>Mr.</title> <username>smithj</username></name>
  <address><street>54 Maple Rive, Santry</street> <country>Irland</country></address>
  <email>smithj@cdt.ie</email>
</user>
```

**RDF (Triple-based approach)**

```
ex:index.html dc:creation-date ex:tem:creation-date "August 16, 1999".
ex:index.html dc:language ex:language "en".
```

*Note: The RDF diagram shows a relationship between resources using properties.*
Reuse existing vocabularies for providing metadata to your resources

General purpose standards and specifications:

- **Dublin Core** for published material (text, images), [http://dublincore.org/documents/dcmi-terms/](http://dublincore.org/documents/dcmi-terms/)
- **FOAF** for people and organisations, [http://xmlns.com/foaf/spec/](http://xmlns.com/foaf/spec/)
- **SKOS** for concept collections, [http://www.w3.org/TR/skos-reference](http://www.w3.org/TR/skos-reference)
- **ADMS** for interoperability assets, [http://www.w3.org/TR/vocab-adms/](http://www.w3.org/TR/vocab-adms/)

Specific standard for datasets:

- **Data Catalog Vocabulary DCAT**, [http://www.w3.org/TR/vocab-dcat/](http://www.w3.org/TR/vocab-dcat/)

Specific usage of DCAT and other vocabularies to support interoperability of data portals across Europe:

Manage your metadata

Ensure good governance through the lifecycle of metadata and of the data it describes.
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.
“A labelling, tagging or coding system used for recording cataloguing information or structuring descriptive records. A metadata schema establishes and defines data elements and the rules governing the use of data elements to describe a resource.”
Designing your metadata schema with RDF Schema (RDFS) – reuse where possible

RDF schema is particularly good in combining terms from different standards and specifications.

Do not re-invent terms that are already defined somewhere else, when designing RDF schemas – reuse terms where possible.

- For example, the DCAT Application Profile for data portals in Europe (DCAT-AP) reuses terms from DCAT, Dublin Core, FOAF, SKOS, ADMS and others.
Example: description of an open dataset with the DCAT-AP

**Description of the Catalogue**

```turtle
:catalog
  a dcat:Catalog ;
  dcat:title "Imaginary Catalog" ;
  rdfs:label "Imaginary Catalog" ;
  foaf:homepage <http://example.org/catalog> ;
  dct:publisher :transparency-office ;
  dcat:dataset :dataset-001 , :dataset-002 , :dataset-003 ;
.
```

**Description of the Dataset**

```turtle
:dataset-001
  a dcat:Dataset ;
  dcat:title "Imaginary dataset" ;
  dcat:keyword "accountability","transparency","payments" ;
  dcat:issued "2011-12-05"^^xsd:date ;
  dcat:modified "2011-12-05"^^xsd:date ;
  dcat:publisher :finance-ministry ;
  dcat:distribution :dataset-001-csv ;
.
```

**Description of the Distribution**

```turtle
:dataset-001-csv
  a dcat:Distribution ;
  dcat:downloadURL <http://www.example.org/files/001.csv> ;
  dcat:title "CSV distribution of imaginary dataset 001" ;
  dcat:mediaType "text/csv" ;
  dcat:byteSize "5120"^^xsd:decimal ;
.
```
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:
  - Example code list - ObjectInCrimeClass (ListPoint)

- **Use concepts identified by a URI** for reference to “things”.
- Example DCAT-AP property:
  - Example taxonomy with terms having a URI - EuroVoc
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 metadata lifecycle

Creating, maintaining, updating, storing, publishing metadata and handling deletion of data.
Creating your 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

**However, 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 resource)?
- How can quality information be included?
Maintaining your metadata

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

• If data does not change, metadata can be relatively stable. Changes (bulk conversions) can take place off-line when needed.

• If data changes frequently (e.g. real-time sensor data), metadata needs to be closely coupled to the data workflow and changes need to be practically instantaneous.
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?
- **Tools and systems** – evolution of storage, bandwidth, mobile...

*Metadata needs to be kept up-to-date to the extent possible, taking into account the available time and budget.*
**Storing your metadata – what are the options?**

*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**.
Handling deletions of data

In many cases, metadata must survive even after deletion of the data it describes.

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 your metadata – what are the options?

• ‘Open’ publication: direct access on URIs
  - This is the option most in line with the vision of Linked Open Data and allows the ‘follow-your-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.

• 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.

See also: [http://www.slideshare.net/OpenDataSupport/licence-your-data-metadata](http://www.slideshare.net/OpenDataSupport/licence-your-data-metadata)
Metadata quality

The quality and completeness of the description metadata of your datasets, directly infect their searchability and reuse.
Metadata quality is about... (1/3)

• The **accuracy** of your metadata - are the characteristics of the resource correctly reflected?
  - *e.g. indicating the right title, the right license, the right publisher enables users to discover resources that they need.*

• The **availability** of your metadata – can the metadata be accessed now and over time into the future?
  - *e.g. making it available for indexing and downloading, and include it in a regular back-up process.*

• The **completeness** of your metadata – are all relevant characteristics of the resource captured (as far as practically and economically feasible and necessary for the application)?
  - *e.g. indicating the license that governs reuse or the format of the distribution enables filters on those aspects.*

See also: http://www.slideshare.net/OpenDataSupport/open-data-quality
Metadata quality is about ... (2/3)

- The **conformance** of your metadata to accepted standards – is the metadata conforming to a specific metadata standard or an Application Profile?
  - *e.g.* following standards and common specifications enables exchange of metadata across portals.

- The **consistency** of your metadata – does the data not contain contradictions?
  - *e.g.* not having multiple and contradictory license statements for the same piece of data

- The **credibility** and **provenance** of your metadata – is the metadata based on trustworthy sources?
  - *e.g.* linking to reference data published and managed by a stable organisation (e.g. the EU Publications Office)
Metadata quality is about ... (3/3)

• The **processability** of the metadata – is the metadata properly machine-readable?
  - *e.g. using references to concepts rather than using free text.*

• The **relevance** of the metadata – does the metadata contain the right amount of information for the task at hand?
  - *e.g. limit the information to optimally serve the users’ needs.*

• The **timeliness** of your metadata – is the metadata corresponding to the actual (current) characteristics of the resource and is it published soon enough?
  - *e.g. indicating the last modification date of the resource enables searches to be filtered on changes after a certain date; making sure the metadata is fresh so that users will see the latest information*
Exchanging metadata

Mapping your metadata to a common metadata vocabulary and exchanging the metadata across platforms.
Homogenising metadata

When exchanged between systems, metadata should be mapped to a common model so that the sender and the recipient share a common understanding on the meaning of the metadata.

- On the **schema level** metadata coming from different sources can be based on different metadata schemas, e.g. DCAT, schema.org, CERIF, own internal model...

- On the **data (value) level**, the metadata properties should be assigned values from different controlled vocabularies or syntaxes, e.g.:
  
  
  - Dates: ISO8601 ("20130101") versus W3C DTF ("2013-01-01")
Example: Homogenising metadata about datasets
The DCAT Application Profile for data portals in Europe

The DCAT-AP can be used as the common model for exchanging metadata with open data platforms across Europe and/or with a data broker (e.g. The Open Data Interoperability Platform - ODIP).

See also: http://joinup.ec.europa.eu/asset/dcat_application_profile/home
Mapping example – data.gov.uk

Scottish Road Accident Statistics

Data about injury road accidents, accident costs, vehicles involved, drivers and riders, drink-drive accidents, drivers breath tested, casualties and international comparisons.

Source agency: Scottish Government
Designation: National Statistics
Language: English
Alternative title: Scottish Road Accident Statistics

Licence
UK Open Government Licence (OGL)

Additional Information

<table>
<thead>
<tr>
<th>Openness score</th>
<th>Scotland</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographic coverage</td>
<td>Travel and Transport</td>
<td></td>
</tr>
<tr>
<td>Temporal coverage</td>
<td>No value</td>
<td></td>
</tr>
<tr>
<td>Date added computed</td>
<td>No value</td>
<td></td>
</tr>
<tr>
<td>Date updated computed</td>
<td>No value</td>
<td></td>
</tr>
</tbody>
</table>

Publisher
Scottish Government

Enquiries:
No details supplied

FOI Contact:
Web:
http://www.whatdotheyknow.com...

Tags
- accident
- health-well-being-and-care
- road
- road-accidents
- road-safety
- roads
- safety
- transport
- transport-accidents-and-casualties
- travel-and-transport

About this dataset
- Added to data.gov.uk: 10/12/2011
- Modified on data.gov.uk: 10/06/2013
- History of changes
- JSON, API and URI for developers

Do more with this data
- Share your app
- Share an idea
- Request new data

Dcat:accessURL, dct:issued, dct:format, dct:description
dct:spatial
dct:temporal
dct:title (Dataset)
dct:language
dct:published
adms:contactPoint
dct:keyword
dct:issued
dct:modified
What can the Open Data Interoperability Platform do?

- **Harvest** metadata from an Open Data portal.
- **Transform** the metadata to RDF.
- **Harmonise** the RDF metadata produced in the previous steps with DCAT-AP.
- **Publish** the description metadata as Linked Open Data.

See also: http://www.slideshare.net/OpenDataSupport/promoting-the-re-use-of-open-data-through-odip
Conclusions

• Metadata provides information on your data and resources. The quality of the metadata directly affects the discoverability and reuse of your resources.

• 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

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

Does your organisation have a minimum set of metadata to be provided together with Open Data?

What would be the main barriers, according to you, for the (re)use of standard controlled vocabularies in your metadata?

Do you have any data and/or metadata governance methodology at the corporate level?
Thank you!
...and now YOUR questions?
References

Slide 6:

Slide 8:
- Dublin City University. Chapter 3: Introduction to XML. http://wiki.eeng.dcu.ie/ee557/g2/326-EE.html
- W3C. RDF Primer. http://www.w3.org/TR/rdf-primer/

Slide 11:
- http://gondolin.rutgers.edu/MIC/text/how/catalog_glossary.htm
- Dublin Core, Example RDF Schema. http://dublincore.org/2012/06/14/dcterms.rdf

Slide 11, 30-32:

Slide 17:

Slide 17-18:
Further reading

Understanding Metadata, NISO.  


Related initiatives


ISA Programme. DCAT Application Profile for European Data Portals, https://joinup.ec.europa.eu/asset/dcat_application_profile/description


The Dublin Core Metadata Initiative, http://dublincore.org/
Be part of our team...

Find us on

Open Data Support
http://www.slideshare.net/OpenDataSupport

Open Data Support
http://goo.gl/y9ZZI

Follow us

@OpenDataSupport

Join us on

http://www.opendatasupport.eu

Contact us

contact@opendatasupport.eu