

## 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

**Training Module 1.4** 

Introduction to metadata management





#### 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

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).





#### Types of metadata

- **Descriptive metadata**, describe a resource for purposes of discovery and identification.
- Structural metadata, e.g. data models and reference data.
- Administrative metadata, provides information to help manage a resource.

In this tutorial we are focusing mainly on descriptive metadata for datasets.

Administrative metadata is also partly covered.





#### Examples of metadata

#### Label

#### Provides metadata on

#### Can







#### Catalogue card



Book



#### Dataset description (DCAT)



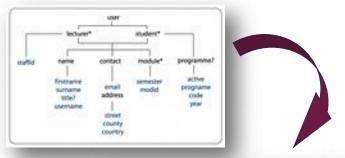


	Day N	ments to	Mind Street,	100
-1	10.1		WW	18.
-	17.5	39	95W	29.
	18.3	59	500	33
-	19.0	65	560	- 18
	18.8		WSW	39
-	18.2	61	. 10W	28
-	12.9	- 61	500	20.



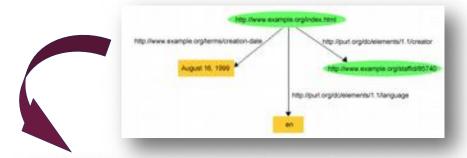
#### Two approaches for providing metadata on the We

#### XML (Tree/container approach)



```
<?sal version="1.0"?>
<!DOOTYPE user SYSTEM "users.dtd">
<ptudent>
   Shame?
     <firstname>Joe</firstname>
      crurname>dmith
     deielabbe, c/elelab
     Cusername>smithj</username>
   4/page?
        <street>54 Maple Rise, Sentry(/street>
        depunty>Dublis
        <country>Ireland</country>
     <email>amithjBdcu.ie</email>
   </enstact>
   cprogramme sotive**true*>
     <code>9608
     cyear>le/year>
   </prepramme>
   <module semester="2">
     cmodid>EE557</modid>
   </module>
   <module semester="1">
     GoodidDEE553K/modid>
   4/module>
</atudent>
```

#### RDF (Triple-based approach)



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



# Managing the metadata of your datasets





#### 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.





#### Metadata schema

"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."

#### XML Schema

```
winesterne enthanes-terripi//www.vd.org/2001/800305enes-mine-terripi/
    targetEssagane-"http://pidi.cog/do/alematia/L:1/" alematFoodeEss
    entrancefrontefactor unquelified.
    Weight about addition?
            workedonimentation and limpetents
                        DOMES 1.1 NOT SITEMA NOT SITEMA THE DATE ! / PURT . DEG/SI/WIRESOLD
                        The Code (trooleddates, edg) The Sebing (ChabongStates, edg) Jame
                          ip. johnstonBukeln. ec. uko. Carl Lagone diagoneBos. cornell. edus 1
                        DC ocenests from the hitsp://gust.org/dc/elemests/1.1/ namespan
                        which parmins mised contect and makes the amirland ettribute at
                        of sundeneura/masteours, however, thus compleating over permit
                         would permit child alessons. All elements are declared as subst
                        means that the deducit type for all elements is doctimulationers
                  NAMES OF TAXABLE PARTY.
            Year Cargon Int. State Opening "State Opening Company 
            entered continue "Attg://www.v0.deg/2001/81/and.and"vi/ancingors/
      A (an) complextype hand" find address?">
            * Our cannot a Choke.
                   wine chooses consider and change and a
                               This is the default type for all of the DC elements. It yests
                                satribute. Too as allowed because signifficant, but satisfies
                                manboursett' are on the sensity tag. This complexitype allows
                        N/MAT SOCIAL PROPERTY AND LOSSES.
                 CONTINUES AND ADMINISTRATION OF
             * CarveonplanContent biass' Tirus's.
                   WIREIDERSCHOOL DE DESETTE BEREIT PROPERTY
                        * DOLLARDSON, NO.
                                     Charlety process/consenter*lay* minfempres*** manifempres***/-
                                n/anrienguezzek
                                couracteritate referred times were extinual to
                        4/MAILTHATCHISCHARD
                  4760 confidences man-
          s/astcomplextypes
          castelement name "aby" type-"himpleliteral" sherous-"(sie"/)
          cantilement name "title" mintitution/out-"any"/s-
          Convelement hasser'sreator' substitutionissup-"any"/b-
          Carrelebant name "subject" substitutionly operany "/b
            CANTED STREET SAME "DESCRIPTION" NAME OF THE PROPERTY OF THE P
             remotenent nemerpublisher" substituturalmuse"eny"/
            CHEROMET NAMES CONTRIBUTORS PROVIDED TO THE PROPERTY NO.
            CAN ADMINIST TAMORADADAY AND ADMINISTRAÇÃO PARTA PARTA
```

```
< humi version+"1.8" encoding+"V/F-6"1>
KIDDCTYNE RDF>
and RDF among notice "Miles! / www.w3.org/2000/01/rdf-schemat"
amins off a "http://www.w3.org/5999/02/22 add systex os#" onto 2ctoring "http://part.org/dc/terms/
 units drawn "http://parl.org/dc/dcare/" units dodn "http://www.wd.org/2004/02/shon/core#"
united date 168pt//seww.w3.org/2002/02/awid-s
   criff Description ref. about > "Milp://purl.org/dc/terms/">
       -ciddema:title emillange "en" > DCMI Metadata Terma - other ciddema:title>
        cildiarms:publisher.nlf.resource="Milgs://perl.org/ds/aboutdoxidDCHE"/:
       colores modified off-detatype="http://www.wit.org/2001/XMLScheme#date">2012-06-
           $4-c/scherma modifiedo-
   - end" Description ethiabout = "MSpc//part.org/dc/terms/Sible" ».
       -crofis:label.com/danger*en*> Tittle-c,/rofis:label
       codic comment and large 'en' a A name given to the resource, a July comment a
       endls is Defined by riff resource "http://purl.org/dc/terms/"/>
       -cichema insued rof: datatype="http://www.wis.org/2001/AMLSchoma#date" > 2008-01
       column modified off datatype="http://www.wil.org/2001/XMLScheina#date">2010-10-
          $$ c/ddwme.modifieds
        and type of resource "Mtg://www.w3.org/1999/02/22-off-syntax-eu#Property"/
       collisions harverson off insource - "Milp://deblincore.org/usage/tones/bistory/#titlet-602"/s
       and among all resource—"http://www.w3.org/2000/01/ndf-scheme#i.html//
        enths submopertyOf off resources "Mttps://pert.org/dc/elements/1.1/title"/>
    chilf-beorgitoes
  - criff Description +th about + "http://purl.org/de/terms/creator"
       -criditative and large 'en' > Creator c
       cofficentment untillange 'en's An entity primarily responsible for making the
          resource, c/rdfs.community
       -cotome description smiling-"en" - Examples of a Creator include a person, an organization, or a
           service, c/dstarmardaucriptio
       cridis/schedule of resources "http://purl.org/dc/terms/"/>
       -cidame issued rdf datatype="http://www.w3.org/2001/XML5chema#date">2008-01
           $4-c/diherry learning
       s.difares modified off-datatype="http://www.wil.org/2001/XMLSchema#date"=2010-10-
          $1 c/ddwmc modifieds
       coll type off resource "http://www.wb.org/1999/02/22-off-syntax-nu#Property"/>
       ditame harmiton of records 'http://dublincore.org/usage/terms/bistory/ecreatort-002'
       enthrouse off-resource-"Mttp://purl.org/dc/teress/Agent"/>
       code submisportyOf off recourse "Mttp://purt.org/dc/elements/1.1/creator"/>
       code authopertyOf off recourses "Mitgs://park.org/dc/terms/contributor"/
        continguis standing only off resources billy / Jambos.com/final/0.1/maker/
  - ordf Description rdf about = "Mttpc//purl.org/dc/terms/subject">
```

RDF Schema







## Reuse existing vocabularies for providing metadat your resources

General purpose standards and specifications:

- Dublin Core for published material (text, images), <u>http://dublincore.org/documents/dcmi-terms/</u>
- FOAF for people and organisations, <a href="http://xmlns.com/foaf/spec/">http://xmlns.com/foaf/spec/</a>
- SKOS for concept collections, <a href="http://www.w3.org/TR/skos-reference">http://www.w3.org/TR/skos-reference</a>
- **ADMS** for interoperability assets, <a href="http://www.w3.org/TR/vocab-adms/">http://www.w3.org/TR/vocab-adms/</a>
  <a href="mailto:Specific standard for datasets">Specific standard for datasets</a>:
- Data Catalog Vocabulary DCAT, <a href="http://www.w3.org/TR/vocab-dcat/">http://www.w3.org/TR/vocab-dcat/</a>
  Specific usage of DCAT and other vocabularies to support interoperability of data portals across Europe:
- DCAT application profile for data portals in Europe,
   http://joinup.ec.europa.eu/asset/dcat\_application\_profile/description

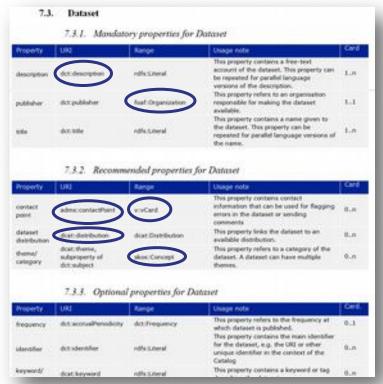


#### Designing your metadata schema with RDF Schem - reuse where possible

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 1

### 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 <a href="http://www.example.org/files/001.csv">http://www.example.org/files/001.csv</a>;
    dct: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 terassigning 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 property

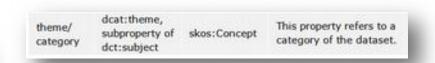
- Use code lists as controlled vocabulary for free text or "string" properties.
- Example DCAT-AP property:



Example code list ObjectInCrimeClass (ListPoint)



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



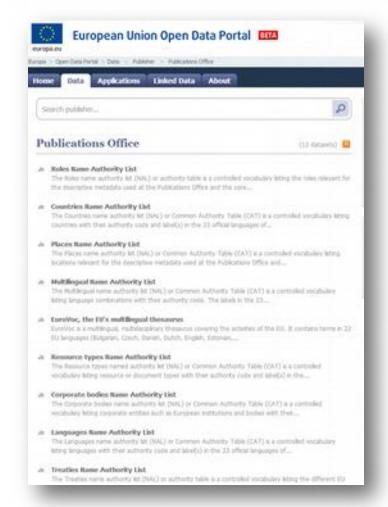
 Example taxonomy with terms having a URI - EuroVoc





#### Example -Publications Office's Named Authority L

- 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 me 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

See also:

http://www.slideshare.net/OpenDataSupport/licence-your-data-metadata







## Metadata quality

The quality and completeness of the description metado your datasets, directly affects their searchability and re





#### 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 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 licence that governs reuse or the format of the distribution enables filters on those aspects.



#### 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. the description of a dataset conforms to the DCAT-AP.
- 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. making the metadata of a dataset available in RDF and/or XML, and not as 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, thus making sure the metadata is fresh so that users will see the latest information.





## Exchanging metadata of datase

Mapping your metadata to a common metadata vocabul such as the DCAT-AP, and exchanging the metadata acreplatforms.





#### Homogenising metadata

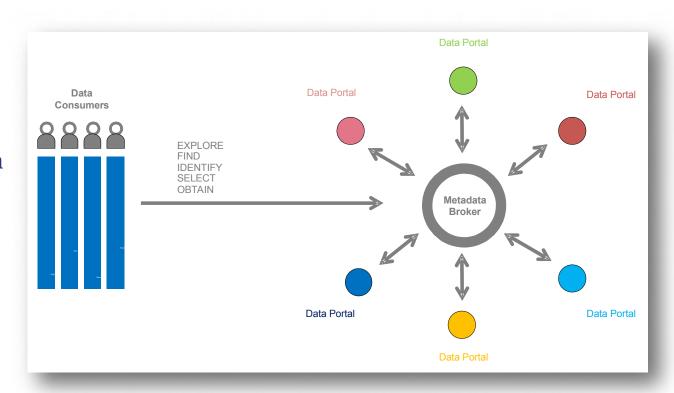
When exchanged between systems, metadata should be mapped to a common model so that the sender and the recipient share a common understanding on the meaning of the metadata.

- 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.:
  - Language: English can be expressed as http://publications.europa.eu/resource/authority/language/ENG or as http://id.loc.gov/vocabulary/iso639-1/en
  - Dates: ISO8601 ("20130101") versus W3C DTF ("2013-01-01")



#### Example: Homogenising metadata about datasets 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



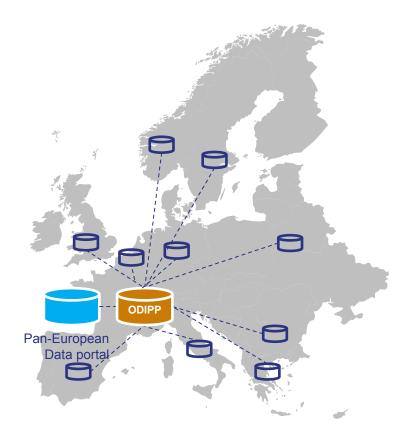


Mapping example - data.gov.uk



#### What can the Open Data Interoperability Platform

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



#### See also:

http://www.slideshare.net/OpenDataSupport/promoting-the-re-useof-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 the 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?

### Take also the online test her





## Thank you! ...and now YOUR question





#### This presentation has been created by Open Data S

#### **Disclaimers**

- 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.
- 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.

Authors:

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





#### References

#### Slide 6, 7:

NISO. Understanding Metadata. http://www.niso.org/publications/press/UnderstandingMetadata.pdf

#### Slide 9:

- 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 12:

- 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

#### Slide 14, 33:

The ISA Programme. DCAT Application Profile for Data Portals in Europe - Final Draft.

https://joinup.ec.europa.eu/asset/dcat application profile/asset release/dcatapplication-profile-data-portals-europe-final-draf

#### Slide 18:

ListPoint. ObjectInCrimeClass. http://www.listpoint.co.uk/CodeList/details/ObjectInCrimeClass/1.2/1

#### Slide 19:

Publications Office. Countries Name Authority List. http://opendata.europa.eu/en/data/dataset/2nM4aG8LdHG6RBMumfkNzQ





#### Further reading



Understanding Metadata, NISO. <a href="http://www.niso.org/publications/press/UnderstandingMetadata.pdf">http://www.niso.org/publications/press/UnderstandingMetadata.pdf</a>



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



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





#### Related initiatives



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



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



ISA Programme. DCAT Application Profile for European Data Portals, 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/generateadms-asset-descriptions-spreadsheet-refine-rdf



The Dublin Core Medatata 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



