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.

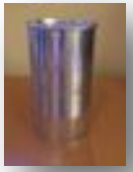
Administrative metadata is also partly covered.

Examples of metadata

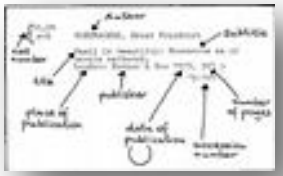
Label

Provides metadata on

Can



Catalogue card

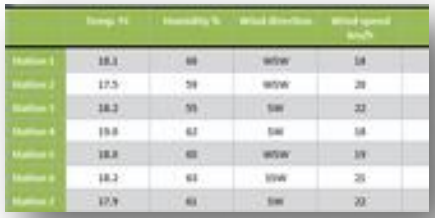


Book

Dataset description (DCAT)

```
:weather1-7 a dcat:Dataset ;
   dct:title "Measurements from weather stations 1-7" ;
   dct:description "Data from seven weather stations
                    showing temparture, 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/CC0> ;
   dcat:downloadURL <http://myweather.com/stations1-7.xlsx>
   .
```

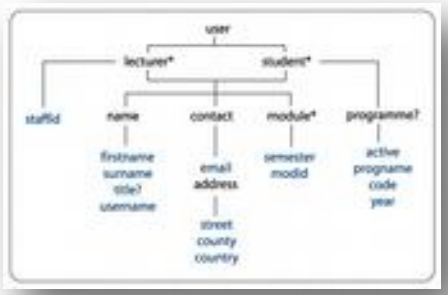
Dataset



CC BY

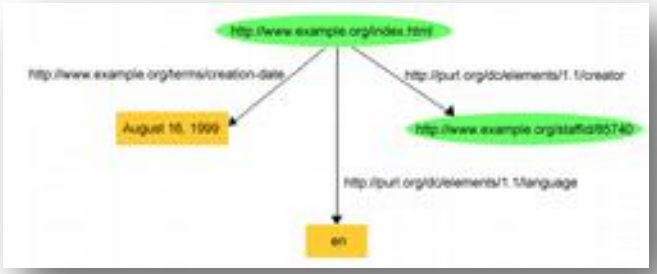
Two approaches for providing metadata on the We

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>56 Maple Rise, Santry</street>
          <county>Dublin</county>
          <country>Ireland</country>
       </address>
       <email>smithj@dcu.ie</email>
    </contact>

    <programme active="true">
       <progname>M.Eng in Electronic Systems</progname>
       <code>MEEK</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" .
```

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

RDF Schema

CC BY

Reuse existing vocabularies for providing metadat your resources

General purpose standards and specifications:

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

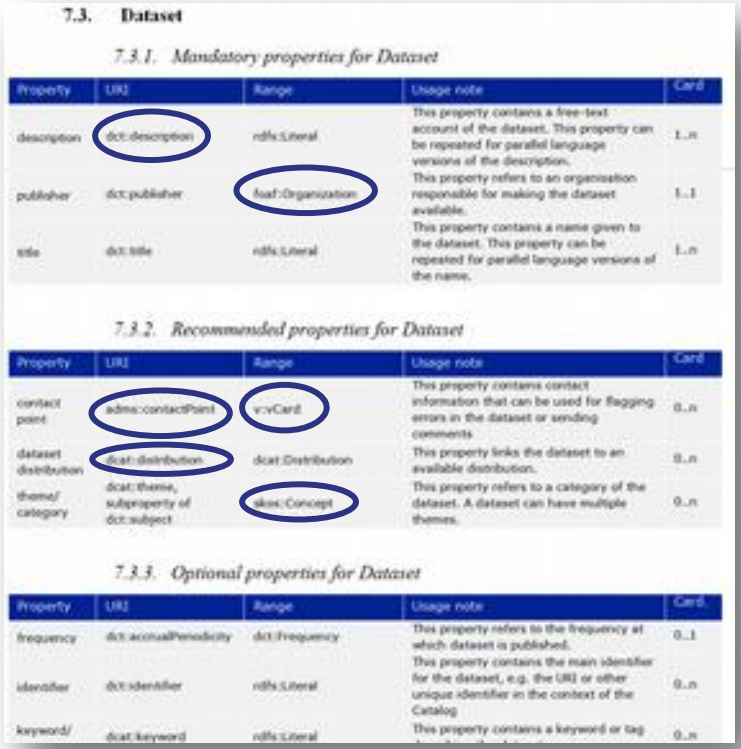
- **DCAT application profile for data portals in Europe**, http://joinup.ec.europa.eu/asset/dcat_application_profile/description

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.



7.3. Dataset

7.3.1. Mandatory properties for Dataset

Property	URI	Range	Usage note	Card.
description	dct:description	rdfs:Literal	This property contains a free-text account of the dataset. This property can be repeated for parallel language versions of the description.	1..n
publisher	dct:publisher	foaf:Organization	This property refers to an organisation responsible for making the dataset available.	1..1
title	dct:title	rdfs:Literal	This property contains a name given to the dataset. This property can be repeated for parallel language versions of the name.	1..n

7.3.2. Recommended properties for Dataset

Property	URI	Range	Usage note	Card.
contact point	adms:contactPoint	v:vCard	This property contains contact information that can be used for flagging errors in the dataset or sending comments.	0..n
dataset distribution	dcat:distribution	dcat:Distribution	This property links the dataset to an available distribution.	0..n
theme/ category	dcat:theme, subproperty of dct:subject	skos:Concept	This property refers to a category of the dataset. A dataset can have multiple themes.	0..n

7.3.3. Optional properties for Dataset

Property	URI	Range	Usage note	Card.
frequency	dct:accrualPeriodicity	dct:Frequency	This property refers to the frequency at which dataset is published.	0..1
identifier	dct:identifier	rdfs:Literal	This property contains the main identifier for the dataset, e.g. the URI or other unique identifier in the context of the Catalog	0..n
keyword/	dcat:keyword	rdfs:Literal	This property contains a keyword or tag	0..n

Example: description of an open dataset with the I

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 ;
    .
```

Controlled vocabularies

Using thesauri, taxonomies and standardised lists of ter
assigning values to metadata properties.

CC BY

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.

CC BY

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:

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

- Example code list - ObjectInCrimeClass (ListPoint)

Code	Meaning
1	Used
2	Taken Without Consent
3	Eliminated
4	Not Involved
5	Involved
6	Crime Weapon
7	Handled / Received
11	Exhibit

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

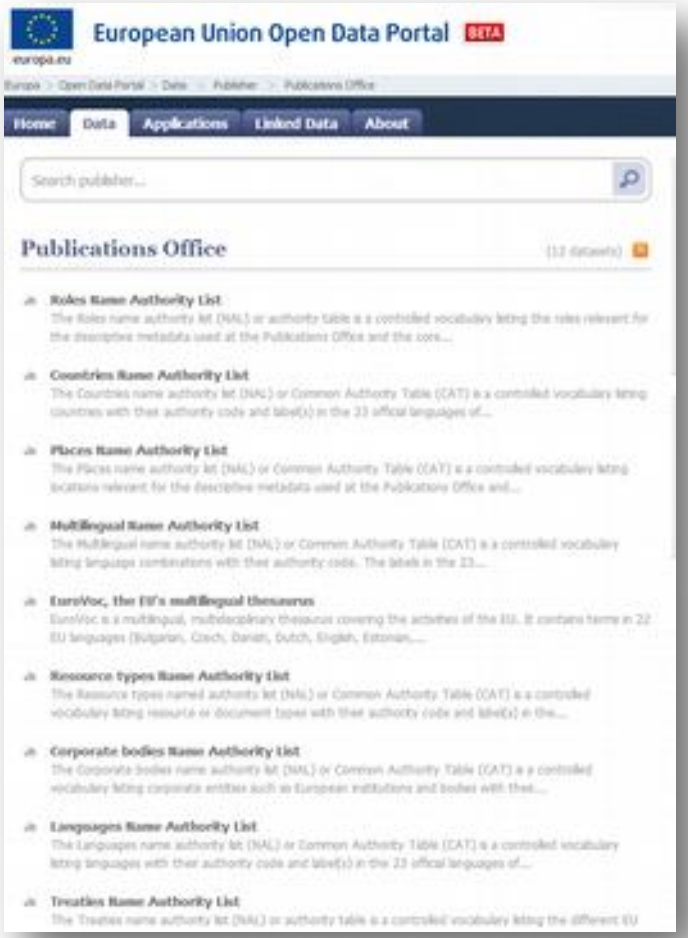
theme/ category	dcat:theme, subproperty of dct:subject	skos:Concept	This property refers to a category of the dataset.

- 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 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.

CC BY

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

Metadata quality

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

CC BY

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.*

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. 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 datas

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

CC BY

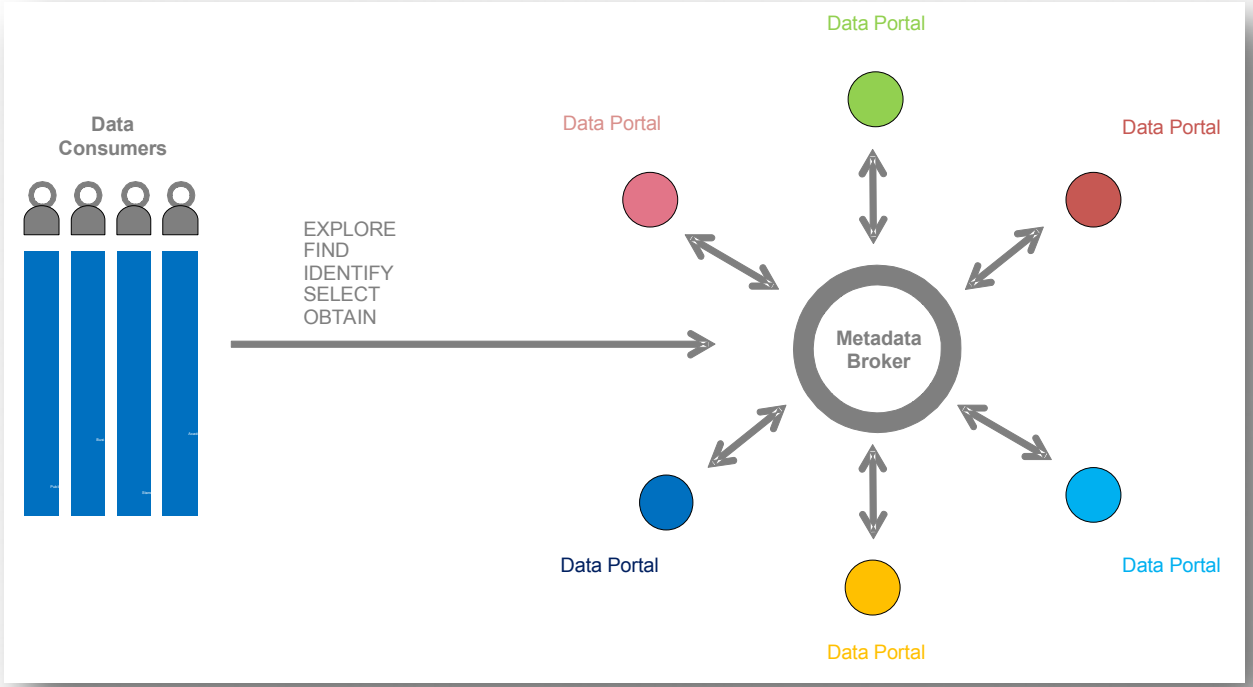
Homogenising metadata

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”)

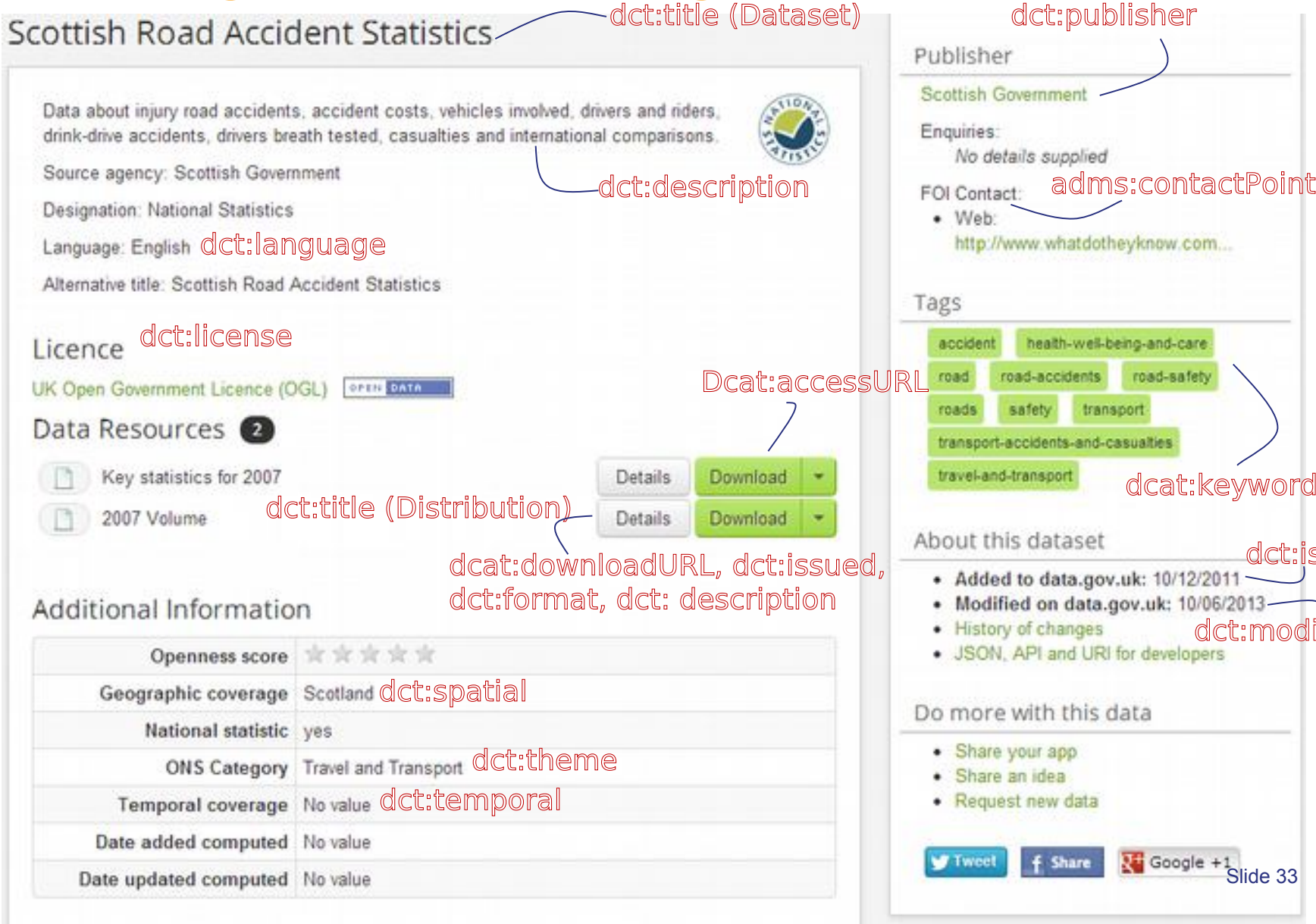
Example: Homogenising metadata about datasets
DCAT Application Profile for data portals in Europe

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).



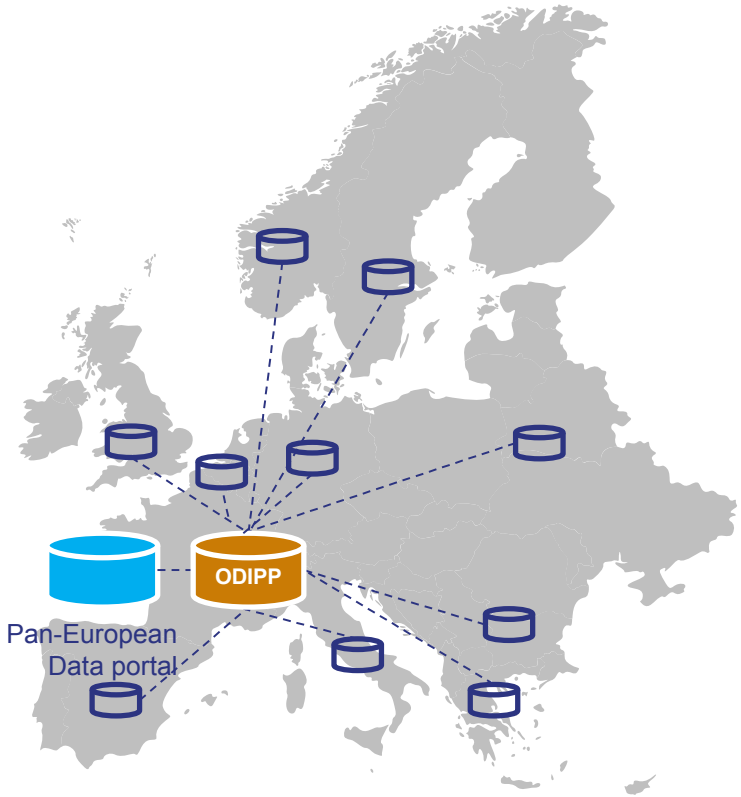
See also:
http://joinup.ec.europa.eu/asset/dcat_application_profile/home

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-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 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.