Harvesting INSPIRE metadata to the pan-European Open Data Portal

Andrea Perego

INSPIRE MIG 2nd F2F Meeting, Arona, 10 April 2014

www.jrc.ec.europa.eu

Serving society
Stimulating innovation
Supporting legislation
The pan-EU Open Data Portal

- The pan-EU Open Data Portal (ODP) is meant to provide a single access point for the cross-border discovery of datasets available from EU data portals.

- This will be achieved by federating existing EU data portals with the pan-EU ODP, and by setting up appropriate harvesting procedures.

- INSPIRE Geoportal to be federated with the pan-EU ODP?
  - Issue raised by Daniele Rizzi (DG CNECT) in his presentation at the ISA SIS WG meeting in Copenhagen (Oct, 2013).
  - A call has been organised on 1 Apr 2014 with the team of the Open Data Support project, who is in charge of setting up the pan-EU ODP federation.
MIG call with Open Data Support project

- Objective: Get precise answers to concerns raised by the MIG on the pan-EU ODP federation and the harvesting criteria
- Most of the discussion was focused on licensing, data sharing, and on how to select “open” datasets
- Many Member States are already publishing INSPIRE metadata also on the national generalistic data portals
- Preliminary conclusions:
  - No direct federation with INSPIRE “geoportals”, unless requested
  - Provide support to interested Member States
  - Avoid duplicating work already carried out in the framework of INSPIRE
DCAT-AP: background

- Purpose: provide a metadata interchange format for data portals operated by EU Member States
- Developed in the framework of the EU Programme “Interoperability Solutions for European Public Administrations” (ISA)
- To be used as a metadata schema for the pan-EU Open Data Portal
- Referred to as a common metadata schema for European data portals in the EU implementation of the G8 Open Data Charter (Oct, 2013)
Based on the W3C Data Catalog vocabulary (DCAT) – now a W3C Recommendation

Defines a minimal set of terms to support metadata interoperability across domains, grouped into mandatory, recommended and optional

It's an “open” metadata profile – i.e., additional terms can be used, without affecting conformity with the specification

Reference representation based on RDF
W3C DCAT
INSPIRE metadata and DCAT-AP

- Why mapping INSPIRE metadata to DCAT-AP?
  - Facilitate cross-sector sharing of INSPIRE metadata
  - Agree upon a common RDF representation
- JRC is completing a discussion paper, proposing an RDF representation of INSPIRE metadata based on DCAT-AP and other relevant vocabularies (DCAT-AP does not provide suitable candidates for all the INSPIRE metadata elements)
- The paper will be submitted to the MIG for review, and to agree upon next steps (if any)
Preliminary results

• Suitable candidates are available for the majority of INSPIRE metadata elements, and the alignment is technically trivial

• The use of HTTP URIs is proposed for the relevant elements, whenever available in the INSPIRE Registry or in the controlledvocabularies recommended in DCAT-AP (e.g., the EU OP's Metadata Registry)

• Two main open issues:
  • Some metadata elements have no suitable candidates in the existing vocabularies
  • The use of global identifiers in INSPIRE metadata is limited and not consistently implemented
The “missing” alignments

INSPIRE metadata elements with no suitable candidate in existing vocabularies

• Spatial resolution
• Coordinate reference system
• Temporal reference system
• Spatial representation type
• Topological consistency
Where global IDs would be beneficial

- Dataset (& services) and, possibly, the corresponding metadata records
- Keywords from controlled vocabularies – including topic categories, etc. (INSPIRE Registry)
- Conditions for access and use & limitations to public access
- Responsible organisation
- Coordinate & temporal reference systems (OGC, EPSG)
(Some) Options for next steps

- About global identifiers & HTTP URIs:
  - Is this really an (urgent) issue?
  - If it is: can we agree on a harmonised way of including global identifiers in the XML-based representation of INSPIRE metadata?

- About missing alignments:
  - Should we exclude the corresponding elements from the DCAT-AP representation of INSPIRE metadata?
  - Should we create our own terms?
Some opportunities

- Contribute the issues identified in the DCAT-AP representation of INSPIRE to existing or upcoming standardisation activities

- A few relevant examples:
  - Work is planned in the framework of the EU ISA Programme to **define extensions to DCAT-AP**
  - Similar work may be also started by W3C, in the framework of the W3C Data Activity
  - The possibility of establishing a joint W3C/OGC working group is currently under discussion. This may also address issues concerning geospatial metadata
More details:

- **W3C Data Catalog vocabulary (DCAT)**
  [http://www.w3.org/TR/vocab-dcat/](http://www.w3.org/TR/vocab-dcat/)

- **ISA DCAT application profile for data portals in Europe (DCAT-AP)**
  [https://joinup.ec.europa.eu/asset/dcat_application_profile/](https://joinup.ec.europa.eu/asset/dcat_application_profile/)

- **EU implementation of the G8 Open Data Charter**

- **Metadata Registry – Publications Office of the EU**

- **W3C Data Activity – *Building the Web of Data***
  [http://www.w3.org/2013/data/](http://www.w3.org/2013/data/)
DCAT-AP: features

- Use of HTTP URIs (*Uniform Resource Identifiers*) for all the main entities defined in the profile
  - Data catalogues
  - Datasets & their distributions
  - Licences
  - Publishers
  - Controlled vocabularies & their terms

- Recommendations for
  - Licences & provenance
  - Controlled vocabularies
Licences & provenance

- **Licences:**
  - Recommended use of “widely recognised licences” (e.g., Creative Commons, Open Data Commons, ISA Open Metadata Licence, EU Public Licence)
  - Mentions work at the Open Data Institute (Open Data Rights Statement vocabulary – ODS) and W3C (Open Data Rights Language – ODRL)

- **Provenance:**
  - Limited support in DCAT / DCAT-AP
  - The use of the W3C PROV vocabulary is recommended
Requirements for controlled vocs

Controlled vocabularies SHOULD:

• Be published under an open licence.

• Be operated and/or maintained by an institution of the European Union, by a recognised standards organisation or another trusted organisation.

• Be properly documented.

• Have labels in multiple languages, ideally in all official languages of the European Union.

• Contain a relatively small number of terms (e.g. 10-25) that are general enough to enable a wide range of resources to be classified.

• Have terms that are identified by URIs with each URI resolving to documentation about the term.

• Have associated persistence and versioning policies.
Recommended controlled vocabularies

For each of the relevant terms in the profile, DCAT-AP recommends the use of specific controlled vocabularies. E.g.:

• EuroVoc
• OP's Metadata Registry Authority Tables (file types, languages, corporate bodies, places)
• ADMS taxonomies (change types, statuses, publisher types, licence types)
• Dublin Core Collection Description Frequency
The Metadata Registry of the EU OP’s

In order to harmonise and standardise the codes and the associated labels used in the Publications Office and on an inter-institutional level in the context of the data exchange between the institutions involved in the legal decision making process, a number of Named Authority Lists (NAL’s) have been defined. These NAL’s are also known as (Common) Authority Tables (CAT’s), controlled vocabularies or value lists.

The following NAL’s are maintained in the Metadata Registry:

1. Released versions

Under IMMC governance

- Corporate bodies
- Countries
- File types
- Interinstitutional procedures
- Languages
- Multilingual
- Resource types
- Roles
- Treaties

Not under IMMC governance

- Currencies
- Places
The Metadata Registry of the EU OP’s

- Give access to a number of registers ("authority tables") concerning language/country codes, formats, corporate bodies, etc.
- Their use is recommended in the DCAT-AP specification
- Already used in the EU Open Data Portal
- It gives also access to EuroVoc and to mappings between EuroVoc and other multilingual thesauri
Thesaurus alignment at the EU OP’s

MDR > EuroVoc

EuroVoc is a multilingual, multidisciplinary thesaurus covering the activities of the EU, the European Parliament in particular. It contains terms in 22 EU languages (Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish), plus Croatian and Serbian.

The following EuroVoc resources (version 4.4) are available for download:

- EuroVoc version (SKOS distribution)
- EuroVoc version (XML distribution)

You will find more information on EuroVoc version 4.4 in the release notes.

The following EuroVoc alignments are available for download:

- Alignment EuroVoc with Agrovoc
- Alignment EuroVoc with Eclas
- Alignment EuroVoc with Gemet

These EuroVoc resources are available as well as datasets on the European Commission Open Data Portal (ODP).

For more information and to browse or search EuroVoc, please visit the EuroVoc website: http://eurovoc.europa.eu
ADMS licence types

**Licence Type**

URL: [http://purl.org/adms/licencetype/1.1](http://purl.org/adms/licencetype/1.1)

rdf:type: [http://www.w3.org/2004/02/skos/core#ConceptScheme](http://www.w3.org/2004/02/skos/core#ConceptScheme)

A vocabulary to classify the conditions and restrictions that are related to the specified licence.

dcterms:rights: European Union

+ skos:hasTopConcept

1. Attribution

   - [http://purl.org/adms/licencetype/Attribution](http://purl.org/adms/licencetype/Attribution)

   skos:prefLabel: "Attribution"@da

   skos:prefLabel: "Zuschreibung"@de

   skos:prefLabel: "Attribution"@en

   skos:prefLabel: "Atribución"@es

   skos:prefLabel: "Attribution"@fr

   skos:prefLabel: "Attribuzione"@it

   skos:prefLabel: "Naamsvermelding"@nl

   skos:prefLabel: "Atrybucja"@pl

   skos:definition: "In general, most licences request to respect author’s credits. Example: Article 5 EULP: “the licensee shall keep intact all copyright notices, … in every distributed copy...” For the most permissive licences, this "Attribution obligation" (combined with a warranty disclaimer) is the main or sole condition for using, copying, performing and (re)distributing the work. Examples: CC by, MIT, BSD, ISA Open Metadata v1.1.”@en

   [http://www.w3.org/2004/02/skos/core#inScheme](http://www.w3.org/2004/02/skos/core#inScheme)

   [http://purl.org/adms/licencetype/1.1](http://purl.org/adms/licencetype/1.1)

2. GrantBack

   - [http://purl.org/adms/licencetype/GrantBack](http://purl.org/adms/licencetype/GrantBack)

   skos:prefLabel: "Grant tilbage"@da

   skos:prefLabel: "Gewähren zurück"@de

   skos:prefLabel: "Grant back"@en
A related vocabulary: ADMS

- Purpose: describe “semantic interoperability assets” (e.g., code lists, controlled vocabularies, thesauri, data models, vocabularies, ontologies)
- Developed in the framework of the EU ISA Programme
- Currently used by several organisations (e.g., EC, W3C)
- Now in W3C space, where it has been defined a profile of the W3C DCAT vocabulary