Aligning INSPIRE Metadata with DCAT-AP
Results, Issues, and Challenges

Andrea Perego
INSPIRE 2014, Aalborg, 18 June 2014

www.jrc.ec.europa.eu

Serving society
Stimulating innovation
Supporting legislation
INSPIRE metadata and DCAT-AP

- **DCAT-AP** (DCAT application profile for data portals in Europe) is meant to 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, which is meant to provide a single access point for the cross-border discovery of datasets available from EU data portals

- Referred to as a common metadata schema for European data portals in the EU implementation of the G8 Open Data Charter (Oct, 2013)
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 has prepared 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 INSPIRE Maintenance & Implementation Group for review, and to agree upon next steps
DCAT-AP: characteristics

• 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

- foaf:Agent
  - dct:publisher
- skos:Concept
  - dct:theme
  - skos:inScheme
  - dct:publisher
- skos:ConceptScheme
  - dcat:themeTaxonomy
- dcat:Catalog
  - dct:title
  - dct:description
  - dct:issued
  - dct:modified
  - dct:language
  - dct:license
  - dct:rights
  - dct:spatial
  - foaf:homepage
- dcat:dataset
- dcat:CatalogRecord
  - dct:title
  - dct:description
  - dct:issued
  - dct:modified
- dcat:Distribution
  - dct:title
  - dct:description
  - dct:issued
  - dct:modified
  - dct:identifier
  - dct:keyword
  - dct:language
  - dct:contactPoint
  - dct:temporal
  - dct:spatial
  - dct:accrualPeriodicity
  - dcat:landingPage
  - dcat:distribution
  - dcat:byteSize
  - dcat:downloadURL
  - dcat:accessURL
  - dcat:mediaType
  - dcat:format
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
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 controlled vocabularies recommended in DCAT-AP (e.g., the Metadata Registry of the Publications Office of the EU).
- 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 (widely used) 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)
Options for next steps

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

- A few relevant examples:
  - Work is being considered in the framework of the EU ISA Programme to define extensions to DCAT-AP
  - Extensions to DCAT may be carried out by W3C, in the framework of the W3C Data Activity (e.g., the Data on the Web Best Practices WG will work on vocabularies concerning data quality and usage)
  - A draft of the joint W3C/OGC Geo Data on the Web WG has been recently published. This WG may also address issues concerning geospatial metadata
**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.
DCAT-AP & 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

- 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
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
   - Corporate bodies
   - Countries
   - Currencies
   - Events (under review)
   - File types
   - Places
   - Interinstitutional procedures
   - Languages
   - Multilingual
   - Resource types
   - Roles
   - Treaties

2. Under preparation
   - Courts
   - Nomenclature of Territorial Units
   - Target audiences
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 Agrivoc
- Alignment EuroVoc with Eclasis
- Alignment EuroVoc with Genet
- Alignment EuroVoc with Inspire
- Alignment EuroVoc with Unibis

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](http://eurovoc.europa.eu)
Interlinking (multilingual) thesauri

• **Purpose**: Interoperable terminologies across communities & organisations
  - Local, regional, national, supra-national levels
  - Different sectors of Public Administrations (PAs)
  - Different scientific disciplines & domains
  - Different languages

• **Practical uses**: Automatically adding terms from given vocabularies (e.g., INSPIRE spatial data themes / EuroVoc) during metadata transformation from one format to another (e.g., INSPIRE / DCAT-AP)

• **Advanced uses**: Support to semantic and cross-language indexing and discovery, also in distributed data sources
EuroVoc and INSPIRE

- The OP has defined *draft* alignments between EuroVoc and
  - INSPIRE spatial data themes
  - INSPIRE feature concepts
- Such alignments will be contributed to the INSPIRE Maintenance and Implementation Group for review and endorsement
- The current version of the alignments is available on the EU Open Data Portal and on the Metadata Registry of the Publications Office of the EU
Thanks!

- **W3C Data Catalog vocabulary (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/

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

- **Metadata Registry – Publications Office of the EU**
  http://publications.europa.eu/mdr/

- **Alignments INSPIRE / EuroVoc – Publications Office of the EU**

- **W3C Data Activity – Building the Web of Data**
  http://www.w3.org/2013/data/

- **W3C Data on the Web Best Practices WG**
  http://www.w3.org/2013/dwbp

- **W3C/OGC Geo Data on the Web WG Charter**
  http://www.w3.org/2014/05/geo-charter