Ontology-based governance of INSPIRE metadata: An implementation scenario

Cristiano Fugazza, Michele Chinosi, and Gianluca Luraschi
{cristiano.fugazza, michele.chinosi, gianluca.luraschi}@jrc.ec.europa.eu

IES - Institute for Environment and Sustainability
Ispra - Italy

http://ies.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/
The role of metadata

- Metadata represents a critical issue in the development of the Web.
- Metadata typically does not play any role in the fruition of data. Nevertheless, it is in general essential to resource retrieval.
- Most of the data provided by spatial data sets and services cannot be efficiently "crawled" by search engines.
- Still, these resources constitute the building blocks of complex applications created by chaining the former according to an abstract work flow.
- Consequently, the availability of metadata, their quality and granularity is a key element for SDIs.
- Finally, the refinement and evolution of metadata need also to be taken into account.
• We explore a novel approach for easing the task of metadata maintenance and improvement.

• We build on the user experience in the data and service retrieval process.

• For doing this, we leverage the main sources for INSPIRE-compliant metadata definition:
  – The INSPIRE Registry;
  – the GEMET Thesaurus.

• We expand the user query according to the relations that link terms in the controlled vocabularies.

• The end user is then given the option to refine her search by to narrowing or broadening the result set.

• Then, the terms that were deemed relevant by the user are proposed to the service provider for inclusion in the metadata set.
The building blocks

• The thesauri were made available according to the Resource Description Framework (RDF) data model.

• The ontology being employed, the Simple Knowledge Organisation System (SKOS), is an effective data schema for representing thesauri.

• SKOS does not involve the learning curve and computational complexity of more expressive ontology definition languages.

• The INSPIRE GeoPortal will be used as a testbed for the new metadata refinement criterion.

• Its search facilities will be examined to understand how to extend its functionalities.

• We will make use of the Business Process Modeling Notation (BPMN) to describe the interaction among the different actors.
The Semantic Web (1)

- RDF decomposes data into “triples” (subject-predicate-object) and can be represented as directed labelled graphs.

N-Triples:

```xml
```

Directed labelled graph (DAG):

```
http://example.org/articleXYZ
  dc:title
  dc:creator
    Cristiano Fugazza
    Michele Chinosi
    Gianluca Luraschi
```

“Ontology-based governance …”
• RDF data can be expressed and serialised in multiple formats.

Turtle:

@prefix dc: <http://purl.org/dc/elements/1.1/>.
<http://.../articleXYZ> dc:title "Ontology-based...";
   dc:creator "Cristiano Fugazza",
             "Michele Chinosi",
             "Gianluca Luraschi".

RDF/XML:

<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/">
   <rdf:Description rdf:about="http://example.org/articleXYZ">
       <dc:title>Ontology-based...</dc:title>
       <dc:creator>Cristiano Fugazza</dc:creator>
       <dc:creator>Michele Chinosi</dc:creator>
       <dc:creator>Gianluca Luraschi</dc:creator>
   </rdf:Description>
</rdf:RDF>
SKOS thesauri

- SKOS allows to structure thesauri according to specificity and relatedness of terms.
• SKOS allows to create thesauri without involving the technicalities of more expressive ontology schema languages.

• This is important for SDI-related thesauri development because domain experts are not ontology experts.

• Spatial data/service providers can then create and interconnect thesauri to be used for service annotation and retrieval.

• Some of the challenges for enabling this are:
  – The identification of user-friendly tools to create/modify thesauri in a collaborative way and export them for storage and usage;
  – the availability of tools for harmonising independent thesauri by distinct thematic communities;
  – the establishment of methodologies for thesauri versioning and governance.
Three main categories of thesauri shall be integrated:

- **General-purpose thesauri**: Provided by governmental bodies or large organisations, typically all-round but not very specific with regard to a given topic, broadly overlapping.

- **Thematic-oriented thesauri**: More promising for an effective annotation of data. Require usable tools (e.g., wiki-based tools). For well bounded formalisms (such as SKOS) form-based interfaces seem to be effective.

- **User-generated thesauri**: Arising from search queries, need to be transparent to the user and involves some heuristics to identify relevant terms. Need to do without a large corpus of “training documents”.
BPMN diagram illustrating the current functioning of the portal:
The technique we are developing comprises a number of distinct phases.

- Query string matching against the thesauri and ranking of results.
- Service metadata records lookup and optional indexing.
- Semantics-aware search refinement on the basis of results.

The query expansion process only involves the text-based pattern inserted by the end user.
Query matching and ranking of results:

- The query string is matched against the natural language descriptions associated with terms in the thesauri. This information is typically divided into a number of distinct fields (*label*, *definition*, *description*).

- The specific field in which the text-based matching occurs will reflect in the ranking of hits (hits in labels will score higher than hits in definitions, etc.).

- For definitions and descriptions, the occurrences of the query string can also be taken into account for ranking purposes.

- The exclusion of either definitions and descriptions by the user could select in order to augment precision in the query matching phase.
Metadata records lookup and indexing:

- We have derived from the textual, multilingual pattern inserted by the user a ranked list of URIs associated with terms in the thesauri that are hosted by the system. The possible usages of this information is twofold:
  - The URI can match those corresponding to mandatory keywords (INSPIRE Themes or the ISO 19119 categorisation of spatial data services).
  - Custom keywords and free-text descriptions may have been related to the URI of terms in the thesauri with a batch procedure.

- In other scenarios, such as that of service brokering, the infrastructure does not have full control over the metadata records of data sets and services. In this case, caching of the associations between URIs and terms is still possible.

- An in-between solution for reducing the run-time computational burden is constituted by the harvesting of metadata descriptions from the catalogues that are brokered, when this functionality is provided by the latter.
Semantics-aware search refinement:

- So far, we only made use of the SKOS-based thesauri to match search pattern in a multilingual context and derive univocal references (i.e., URIs).

- We did not leverage the capability of thesauri to structure terms according to specificity and relatedness.

- The next step of the retrieval process is to allow the end user to refine the search by choosing, for each of the terms retrieved in the first phase, whether to extend or shrink the results set.

- In order to abstract from the technicalities of SKOS, we are going to define only four generic categories of query expansion for the user to choose from:
  - More general terms;
  - more specific terms;
  - corresponding terms;
  - related terms.
Case study: The INSPIRE Geoportal (6)

- BPMN diagram illustrating the semantics-aware extension to the Geoportal search functionalities:
To improve the quality of metadata, it is necessary to provide means that are straightforward and as less time consuming as possible.

From the search work flow we get:

- The identifiers of services accessed by the user.
- The URIs of terms that were selected by the user during the search.

For terms that are not currently included in the keywords associated with services we create or update a database record of the following form:

\[<r, t, c>\]

\[r = \text{the identifier for the resource in the resource set } R\]
\[t = \text{the identifier for the term (a URI) in the thesaurus } T\]
\[c = \text{a counter initially set to 1}\]
• Once the value stored by the counter exceeds a given threshold value $\tau$, the notification process is triggered.

• In the periodical email that is sent to the metadata maintainer, the term will be proposed for inclusion in the metadata set.

• The set of terms $T'$ that are proposed is the following:

$$T' = \{ t : <r,t,c> \land \frac{c}{\{<r',t,c'>\}} > \tau \}$$

$r = \text{the identifier for the resource in the resource set } R$
$r' = \text{the identifier for generic resources in the resource set } R$
$t = \text{the identified for the term (a URI) in the thesaurus } T$
$c = \text{the counter associated with resource } r$
$c' = \text{the counter associated with resources } r'$
Thank you for the attention.

Time for questions.