A multilingual/multicultural semantic-based approach to improve Data Sharing in a SDI

Monica De Martino, Riccardo Albertoni
IMATI-CNR-GE
Institute of Applied Mathematics and Information Technologies, Dept. Genoa, Italian National Council of Research, Italy
Overview

• Background -
  • NatureSDIPlus Project
  • Why a thesaurus? Where are we going to use it?

• Common thesaurus framework for Nature Conservation
  • Why considering existing terminological resources?
  • Design principles
  • Implementation exploiting SKOS and Linked data
Background (1)

• NatureSDIPlus (http://www.nature-sdi.eu/) EU founded project, eContentplus programme
  • it aims at establishing a SDI for Nature Conservation.
  • it supports the implementation of the INSPIRE Directive on the data themes:
    • Protect Site (Annex I); Biogeographical region; Habitat and biotopes and species distribution (Annex III).
NatureSDIPlus expected results

- a metadata profile for the specific thematic community covered by the project and based on ISO 19115/119 standards.
- a Common data model compliant with INSPIRE specifications to improve the harmonization of available datasets made available through designed web services.

Multicultural & Multilingual Thesaurus of common terms in Nature Conservation to be exploited as Keywords in the metadata, Categorical attribute in the Data Model, Search Keyword in the web portal
Define a brand new thesaurus?
Don’t reinvent the wheel!

1. different communities with a large spectrum of competencies are involved in the Nature Conservation;
2. many terminologies have been already developed and adopted on these competencies;
3. more that one terminology can be available for a given competency;
4. terminologies adopted have often a national origin, so they are not uniform in all the European countries and often even stakeholders from the same country can adopt different terminology in everyday practice.
Common thesaurus framework

Integrating well known existing thesauri or classification.

Framework Design Requirements

- Modularity:
  - Each new thesaurus can be added as a new module in the framework

- Openness:
  - Each terminology/thesaurus should be easily extendable

- Interlinking:
  - Interlinking among the terms and concepts of different available thesauri in order to harmonize their usage

- Exploitability:
  - Framework thesauri encoded in a standard and flexible format to encourage the adoption and its enrichment from third parties user and system
Common thesaurus framework

basic idea
Implementation
Methodology

- Action 1: Sources selection
- Action 2: Thesaurus encoding (SKOS)
- Action 3: Thesaurus publication with Linked Data
- Action 4: Thesauri interlinking
- Action 5: Thesaurus Validation
### Action 1: Sources Selection

| Thesaurus Classification | Sources | Number of terms | Language | Data Theme covered          |
|--------------------------|---------|-----------------|----------|-----------------------------|
| EARTH                    | CNR     | 14340           | en, it   | All (General level)         |
| GEMET                    | EEA     | About 6562      | (en, es, nl, el, pt, it, fr, da, fi, sv, de, ..) | ALL (General Level) |
| IUCN classification      | IUCN    | 8               | en       | Protect Site                |
| Habitat types            | EUNIS DB| 5431            | en       | Habitat & Biotopes          |
| Nature 2000              | EUNIS DB| 183447          | Latin    | Species Distribution        |
| EUNIS Species            | EUNIS DB| 12              | en       | Biogeographical Region      |
| Main threats to biodiversity by biogeographic region | EEA | About 68 | en | Biogeographical Region |
| Digital map on EU Ecological Regions | DMEER | 68 | en | Biogeographical Region |
Action 2: Thesaurus encoding (SKOS)

• SKOS—Simple Knowledge Organization System
  • W3C Recommendation, 18 August 2009
• Encoding in a model to be available also in Semantic Web environment
• Examples of Thesauri already available in SKOS
  • GEMET (http://www.eionet.europa.eu/gemet),
  • AGROVOC (ftp://ftp.fao.org/gi/gil/gilws/aims/kos/agrovoc_formats/owl/)
What Skos Model is for?

SKOS

Animal

BT

Cat

http:\xyz\id1

skos:prefLabel

animal@en

Animale\@it

http:\xyz\id2

skos:prefLabel

skos:broader

cat@en
tomcat@en

skos:altLabel
Existing Thesauri Extension

Existing Thesauri Mapping

http:\xyz\id1
http:\xyz\id2

http:\yyy\id1
http:\yyy\id2

É.

http:\xxx\id1
http:\xxx\id2

animal@en
kichen@en

dog@en

farm animal@en

cat@en

guard dog@en

tomcat@en
Action 3: Thesaurus Publication with Linked Data

- Linked Data a smart way to publish Thesauri into the web
  - Makes the thesaurus available both for Humans and Machines
- It makes easier to meet the requirements of:
  - **Modularity,**
  - **Openness** for further thesaurus extension
  - **Ability of interconnect** the thesaurus to other thesaurus
  - **Exploitability** of the thesaurus from other third party system
Action 4: Thesauri Interlinking
Action 5: Thesaurus Validation

A questionnaire distributed to the NatureSDIPlus Consortium to check

- if the Consortium agrees with the Expert Team in relation with the selected resources for the four data theme domains;
- if the Common Thesaurus Framework is useful in the multilingual information search and retrieval.
- if the Common Thesaurus Framework is useful in the metadata compilation;

Results are quite encouraging!!!
Which of the included thesaurus/classification can be used for which NatureSDIplus themes?

| Thesaurus/Classification | Protect Site | Biogeographical regions | Habitat & Biotopes | Species Distribution | None |
|--------------------------|--------------|-------------------------|-------------------|---------------------|------|
| EARTH                    | 30,8         | 38,5                    | 61,5              | 53,8                | 23,1 |
| GEMET                    | 53,8         | 53,8                    | 69,2              | 61,5                | 23,1 |
| IUCN Classification      | 84,6         | 0,0                     | 0,0               | 23,1                | 7,7  |
| Habitat Types & NATURA 2000 A I | 38,5     | 30,8                    | 84,6              | 61,5                | 7,7  |
| EUNIS Species            | 15,4         | 23,1                    | 46,2              | 84,6                | 7,7  |
| Main threats to Biodiversity by European Biological Regions / Digital Map on Ecological Regions (DMEER) | 7,7 | 84,6 | 7,7 | 0,0 | 15,4 |
Do included thesaurus/classification help for data search in the NATURE-SDIplus geoportal?

|                                | No  | Yes  | I don't Know |
|--------------------------------|-----|------|--------------|
| **EARTH**                      | 15,4| 30,8 | 53,8         |
| **GEMET**                      | 0,0 | 69,2 | 30,8         |
| **IUCN Classification**        | 0,0 | 84,6 | 15,4         |
| **Habitat Types & NATURA 2000 A I** | 0,0 | 92,3 | 7,7          |
| **EUNIS Species**              | 7,7 | 76,9 | 15,4         |
| **Main threats to Biodiversity by European Biological Regions / Digital Map on Ecological Regions (DMEER)** | 0,0 | 53,8 | 46,2         |
Conclusion

We have provided an open, modular, extensible, and exploitable thesaurus framework exploiting SKOS and Linked Data Best Practice.

If you have suggestions, you want to extend, exploiting our work.

Please visit

http://linkeddata.ge.imati.cnr.it:2020/

And/or write to

riccardo.albertoni@ge.imati.cnr.it
monica.demartino@ge.imati.cnr.it