+

Networked Knowledge Organization Systems and Services The 9th European NKOS Workshop at the 14th ECDL Conference, Glasgow, Scotland 10 September 2010



SKOS and semantic web best practice to access terminological resources: NatureSDIPlus and CHRONIOUS hand-on experience

Riccardo Albertoni, <u>Albertoni@ge.imati.cnr.it</u> Monica De Martino, Franca Giannini, IMATI-CNR-GE, Italy

+ Goals of this presentation

- To share hand-on experience we got working in two European Projects (NatureSDIPlus and CHRONIOUS)
 - Motivations which brought us deploying KOS in the projects
 - SKOS + linked data in NatureSDIPlus
 - SKOS + OWL Ontologies in CHRONIOUS
 - Common abstract pipeline to set up and exploit KOS,
 - Deployments-instantiation of such a pipeline according to constraints arising in NatureSDIPlus and CHRONIOUS projects

To provide

- Hand-on recipes: hopefully, you can adopt, adapt, and enhance our solutions
- bases for a critical discussion
 - Suggestions from the audience are welcome

NatureSDIPlus

- ECP-2007-GEO-317007 <u>http://www.nature-sdi.eu/</u>
 - Best Practice Network aimed at establish a Spatial Data Infrastructure (SDI) for Nature Conservation
 - October 2008-2011,(30 months)
 - to enable and improve the harmonisation of national datasets on nature conservation. The considered data themes are: Protect Site (Annex I); geogeographical region, Habitat and biotopes and species distribution (Annex III).

We were leading the TASK defining a terminology\thesaurus as common base for Metadata, keyword search

CHRONIOUS

- FP7-ICT-2007-1-216461, <u>http://www.chronious.eu/</u>
 - An Open, Ubiquitous and Adaptive Chronic Disease Management Platform for COPD and CKD
 - February 2008- 2012, (48 months)
 - to define a European framework for a generic health status monitoring platform addressing people with chronic health conditions. This will be achieved by developing an intelligent, ubiquitous and adaptive chronic disease platform to be used by both patients and clinicians

We were involved in the Thesaurus-Ontology module supporting the search for scientific literature pertaining to the COP and CK deseases

Why KOS in NatureSDIPlus?

╋

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; (but still different formats and models)
- 3. more than 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 even stakeholders from the same country can adopt different terminologies in the everyday practice.

+ Common thesaurus framework

Integrating well known existing thesauri or classifications.

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 is allowed in order to harmonize terminologies
- Exploitability: Framework thesauri encoded in a standard and flexible format to encourage the adoption and its enrichment from third parties user and system



+ SKOS





Common thesaurus: Integrating well known existing thesauri or classifications.

- SKOS/RDF as Common thesaurus format supporting the multilingualism
- SKOS/RDF + Linked data best practices paving the way for
 - 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 Current state



Why KOS in CHRONIOUS?

╋

Terminology to index scientific literature

- MeSH is a well known controlled vocabulary used for indexing articles from MEDLINE/PubMed
 - But it isn't enough specialized to deeply cover COPD and CKD
- Formal Ontologies have been defined to deepen these diseases
 - MloC (middle layer), COPD and CKD ontologiesprovided by IFOMIS
- However MeSH is still required in Chronious
 - The search is not always made at the same level of granularity, often keywords search can be done moving back and forward from coarse to very diseasespecialized concepts
 - Multilingual support, some "certified" translation are available for example in it, pt, es
 - Terminological de facto standard, Clinicians expect it is included
- How to combine ontologies and MESH in CHRONIOUS ?
 - A Skossyfied version of MeSH and we used RDF as a kind of lingua franca

+ CHRONIOUS's KOS

MeSH 2010 Skossyfied

Skossyfied MESH Translations in Italian, Portuguese, Spanish

SKOS- RDF -URI

IFOMIS's Specialized Ontologies In OWL Mapping between Skossyfied MeSH and Ontologies

A pipeline to set up KOS

╋

Pipeline wrt projects



╋

Resource Selection

+ Which resources?

- In NatureSDIPlus,
 - How to manage feedbacks from experts with limited time and economic resources?
 - we have more than 30 partners involved (with Multiple competencies/fields of expertise)

SUGGESTION:

Questionnaire to partners and partners' friends by surveymonkey.com

Restricted group of experts for revise the feedbacks

• Are the resources suggested available in electronic form?

Selection has been rediscussed with all partners by second electronic questionnaire

Copyright?

• Extremely tricky: It was extremely hard to find out

- Who is the owner of the data..
- If we could use and republish data..
- Under which restrictions ..

Example in NatureSDIPLUS: We asked to who distributed the data and the owner and then we get a mail saving go ahead III E Suggestions: • to deal with copyright issue since the earliest phase of resource selections \blacksquare N •Selecting resources that cannot be exploited as you need might d jeopardize your project efforts •Take a look at initiative that have been establish in the meanwhile, but if Exa had to face this problem now I would start from but http://www.opendatacommons.org/quide/ p p •http://www.slideshare.net/jordanhatcher/linked-data-licensingintroduction-isemantics-2010

+ Translation into SVOS

Translation into SKOS

What we have used ...

D2R Server, http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

- To map relational DB into RDF vocabularies
- To publish vocabularies as linked data
- To dump the data as RDF
- Open source from Freie Universität Berlin
- Very simple, if you know SQL, (mySQL), you have just to learn D2RQ the mapping language

Consideration:

provides a p SPARQL and

Jena, <u>http://je</u> I would recommend such a bunch of technology at least as is a Java frar starting point

•Tool and framework available for free

open source •Very limited technological knowledge is required Programme. •Basic semantic web\linked data principle •JAVA – MySQL

+ Where we have used what ..

Project	Resources	SKOSifycation
NatureSDIPlus	Excel, Relational Data base	 Importing in MySQL, extraction of a simplified data view D2R server
CHRONIOUS	XML MESH 2010 DUMP Italian MESH Translation	 •Conversion in MYSQL •Importing in MySQL, •extraction of a simplified data view •D2R server •DUMP to RDF
	Spanish and Portuguese MESH Translations	Ad hoc program developed with JAVA and JENA to read a file and convert the info into SKOS\RDF

+ Publication/Access

Project	Kind of Access
NatureSDIPlus	Linked data
NatureSDIPlus /	Web Services similar to SKOS GEMET API
CHRONIOUS	Ad-hoc API

Suggestion:

According to our experience linked data is very good for sharing your resources with third parties enabling them to extend your resources

However, harvesting can be very costly So It is very useful to provide also updated dump copies of your resources

http://www.mdweb-project.org/web

We developed ad-Hoc API that have been included in the CHRONIOUS Architecture

(Inter)Linking

╋

Strategies for Interlinking

- Exploitation of domain experts:
 - The interlinking can be defined manually by the domain experts.
 - Huge efforts, very tricky to reach a consensus especially when a large group of experts are involved.
 - The process can result in a high quality mapping, but only if domain experts are very willing and knowledgeable.
- Exploitation of a priori knowledge:
 - Very often KOS have been created by common origins or they have been built including part of other pre-existing resources.
 - Knowledge about these interrelations can be crucial to link different KOS
 - generally accepted naming schemata, for instance, DOI for libraries, habitat classification as NATURA 2000 A I
 - If the link source and the link target data sets already support one of these identification schemas, the implicit relationships between entities in data sets can easily be made explicit.
- Exploitation of automatic tools:
 - The idea behind these tools is to compare concepts belonging to distinct KOS assessing their similarity, and then they link the concepts whose similarity is higher than a given threshold.
 - SILK, discovering relationships between data items within different Linked Data sources http://www4.wiwiss.fu-berlin.de/bizer/silk/

Common thesaurus framework Interlinking



Interlinking

- Exploitation of domain experts (EARTh-BiogeographicalRegions skos:relatedMatch)
 - We asked directly to the EARTh team to figure out the connections between EARTh and BiogeographicalREgions
 - It worked because their valuable expertise on EARTh and the limited number of concept in BiogeographicalRegions (80 concepts)
- Exploitation of a priori knowledge (EARTh-GEMET, skos:exactMatch)
 - EARTh is an extension of GEMET, when a concept come from GEMET they internally kept the GEMET identifier
 - E.g., Wood (ID 30510) has GEMETID 9349 within EARTH then GEMET URI

(http://www.eionet.europa.eu/gemet/concept?cp=9349)

+ Interlinking

Exploitation of automatic tools (EUNIS Habitat and Species skos:relatedMatch)

Species are easily **identifiable** in the Habitat title and description !!!! Example of HABITAT : Low energy litoral rock

skos:definition Sheltered to extremely sheltered rocky shores with very weak to weak tidal streams are typically characterized by a dense cover of fucoid seaweeds which form distinct zones (the wrack [**Pelvetia canaliculata**] on the upper shore

on the

We didn't use SILK, we defined Ad hoc procedure in JAVA +JENA :

For each HABITAT Y Extract from Habitat Title and Description A={a1, a2, a3,.., an) For each X in A then URI(X) skos:relatedMatch URI (Y) and URI (Y) skos:relatedMatch URI (X)

CHRONIOUS: Mapping between MeSH and Ontologies



Example:



Mapping between MeSH and Ontologies



- Obtained by a two steps process
 - First step: automatic syntatic comparison between ontologies class labels and MeSH terms
 - mesh:mapToEquivalent are created
 - Second step: manual check
 - To delete wrong mapping
 - Concept whose terms have same syntax but different semantics
 - To specialize the mapping if required in
 - mesh:mapToNarrower
 - mesh:mapToBroader

+ Advertising

Void: **Vocabulary of Interlinked Datasets**

- Void provides metadata for your resources, It makes available info about
 - License, data dumps, sparqlEndPoints, Interlinked dataset, exploited RDF vocabulary, example of Resources, homepage
- DETAILED INSTRUCTIONS
 - http://vocab.deri.ie/void
 - http://semanticweb.org/wiki/VoiD
- **EDITOR:**
 - ve² <u>- the voiD editor</u>, <u>http://ld2sd.deri.org/ve2</u>

Let's provide a Void to Earth

First suggestion: Define an URI for each of your resources

<http://purl.org/NET/Earth> rdf:type void:Dataset ;

- You need a stable URI, so I suggest to exploit some service to try to have the URI under your control
 - E.g. PURL to set up an URI (http://purl.com)
- THINK twice before using the URL of the KOS web page as URI
 - What happen if this URL changes?
 - Example 1: you use your company\Institute server and eventually the company\institute changes its name
 - From http://mycompany/mydataset to <u>http://??/mydataset</u>

Let's provide a Void to Earth

Second Suggestion: Beware about Functional Inverse Properties (e.g., foaf:homepage)

<http://purl.org/NET/Earth> foaf:homepage <http://ekolab.iia.cnr.it/earth_eng.htm>;

Pay attention!!!

foaf:homepage is an Inverse Functional Property



+ Let's provide a Void to Earth

<http://purl.org/NET/Earth> rdf:type void:Dataset; foaf:homepage <http://ekolab.iia.cnr.it/earth eng.htm>; dcterms:title "EARTh"; dcterms: description "Environmental Applications Reference THesaurus"; dcterms:publisher <http://dblp.l3s.de/d2r/resource/authors/Riccardo_Albertoni>; dcterms:license <http://purl.org/NET/EARTHlicence>; **void:sparqlEndpoint** < http://linkeddata.ge.imati.cnr.it:2020/sparql>; void:dataDump <http://purl.oclc.org/net/DumpEarthRDF>; **void:vocabulary** <http://www.w3.org/2004/02/skos/core#>; **void:vocabulary** <http://www.w3.org/1999/02/22-rdf-syntax-ns#>; void:exampleResource <http://linkeddata.ge.imati.cnr.it:2020/resource/EARTh/100000>;

void:exampleResource

<http://linkeddata.ge.imati.cnr.it:2020/resource/EARTh/13040>; dcterms:subject <http://dbpedia.org/resource/Natural_environment>; dcterms:subject <http://dbpedia.org/resource/Thesaurus>; void:subset :myDS-DS1 ; # EARTh has also a subset :myDS-DS1 void:subset :myDS-DS2 . # EARTh has also a subset :myDS-DS2

Let's provide a Void to Earth

that are linked to GEMET

:DS1 rdf:type void:Dataset;

foaf:homepage <http://eionet.europa.eu/gemet> ;

dcterms:title "GEMET";

dcterms:description "GEneral Multilingual Environmental Thesaurus "; **void:exampleResource**

<http://www.eionet.europa.eu/gemet/concept?cp=8344>.

```
:myDS-DS1
    rdf:type void:Linkset;
    void:linkPredicate <http://www.w3.org/2004/02/skos/core#exactMatch>;
    void:target <http://purl.org/NET/Earth>;
    void:target :DS1.
```

+ Good .. What once we have Written a VOID description

Publish it

- Ve2 provides a support to notify the VOID to <u>Sindice</u>, the <u>RKB voiD</u> <u>store</u>, the <u>Talis voiD store</u>, and to <u>PingtheSemanticWeb.com</u>.
- Publish as RDF in your Web site and notify by your own where to harvest the VOID description of your data

Search your dataset

- By Sindice, (sindice.com)
 - semantic web index to search RDF fragments
 - Different query form: keywords, relations (* foaf:knows C), advance query (AND, OR, ..)

Searching GEMET on SINDICE



Other SINDICE queries

http://sindice.com/

- Give me the list of dataset for which Riccardo Albertoni is Publisher
 - * <http://purl.org/dc/terms/publisher> <http://dblp.l3s.de/d2r/resource/authors/Riccardo_Albertoni>
- Give Me the list of dataset VOID whose Riccardo Albertoni is publish
 - (* <http://purl.org/dc/terms/publisher> <http://dblp.l3s.de/d2r/resource/authors/Riccardo_Albertoni>) AND (* <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://rdfs.org/ns/void#Dataset>)
- If you ask for GEMET you get also EARTH
 - * <http://xmlns.com/foaf/0.1/homepage><http://eionet.europa.eu/gemet>
- Give me all RDF fragment pertaining to <u>http://www.eea.europa.eu/data-and-maps/data/digital-map-of-european-ecological-regions</u>
 - * <http://xmlns.com/foaf/0.1/homepage> <http://www.eea.europa.eu/dataand-maps/data/digital-map-of-european-ecological-regions>

+ Conclusion - Discussion

For further information & questions please write to Riccardo.Albertoni@ge.imati.cnr.it