EARTh: an Environmental Application Reference Thesaurus in the Linked Open Data Cloud

R. Albertoni[a,b,*], M. De Martino[b], S. Di Franco [c], V. De Santis [c] and P. Plini[c]
[a]*Ontology Engineering Group. Dpto. de Inteligencia Artificial Facultad de Informática, Universidad Politécnica de Madrid 28660, Boadilla del Monte, Madrid, Spain, ralbertoni@fi.upm.es*
[b]*CNR-IMATI, Via De Marini, 6, 16149 Genova, Italy, demartino@ge.imati.cnr.it*
[c]*CNR-,IIA EKOLab - Environmental Knowledge Organization Laboratory, Area della Ricerca di Roma 1, Via Salaria Km 29,300 C.P. 10,I-00016 Monterotondo stazione RM, plini@iia.cnr.it*

Abstract. The paper aims at providing a description of EARTh, the Environmental Application Reference Thesaurus. EARTh represents a common general terminology for the environment, which has been published as a SKOS dataset in the Linked Open Data cloud. It promises to become a core tool for indexing and discovery environmental resources by refining and extending GEMET, which is considered the de facto standard when speaking of general-purpose thesaurus for the environmental domain in Europe. The paper illustrates the main key characteristics of EARTh as a guide to its usage. It clarifies (i) the methodology adopted to define the EARTh content; (ii) the design and technological choices made publishing EARTh as Linked Data; (iii) the information pertaining to its access and maintenance. Descriptions of EARTh applications and future relevance are also highlighted.

Keywords: SKOS, Linked Data, EARTh, Thesaurus, Environment

1. Introduction

Although different directives (e.g. INSPIRE [1]) and policy communications (e.g. SEIS [2]) have been launched at European-scale with the objective of improving the management of heterogeneous environmental data sources, an effective sharing of these resources is still desiderata due to the intrinsic multicultural and multilingual nature of the environmental domain.

Thesauri are widely employed as common ground enabling communication among the different communities working in environment-related domains: they allow users to share and agree upon scientific/technical terms in the target domain and to express them in multiple languages. In the recent years several controlled vocabularies (thesauri) have been deployed but as they have been created embodying different points of view, a networked access to heterogeneous environmental data sources requires interoperability of different controlled vocabularies [3]. The Linked Data publishing paradigm [4] jointly with Simple Knowledge Organization System [5] provides a promising framework to represent and publish distinct thesauri and their interlinks as a whole.

This paper presents EARTh, the Environmental Application Reference Thesaurus that takes advantage of this framework providing a SKOS dataset recently included in the Linked Open Data (LOD) Cloud. Compared to other environmental thesauri available as linked data as AGROVOC[1], EUNIS[2], Geological Survey of Austria (GBA) Thesaurus[3], EARTh provides a more general purpose and thematically neutral terminological support. Compared to the GEneral

[1] http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[2] http://eunis.eea.europa.eu/
[3] http://test.ckan.org/it/dataset/geological-survey-of-austria-thesaurus

Multilingual Environmental Thesaurus (GEMET)[4], namely the de facto general purpose thesaurus standard, EARTh provides a minor multilingual support, but it extends GEMET with more than 9000 concepts and revises the GEMET concept hierarchy. Being one of the largest general purpose and structured environmental terminological resources available in the LOD cloud, EARTH is expected in close future to become a good linking point for other terminological resources dealing with environmental topics. Anyway, EARTh already includes a linkset with more than 4000 equivalence between EARTh and GEMET concepts, enabling the inter-thesaurus navigation from EARTh to GEMET, the traditional thesaurus-based indexing of digital resources, as well as the use of digital resources across multi-thesauri applications and platforms.

The remainder of this paper is organized as follows: Section 2 describes EARTh in terms of its content, the methodology followed and extension of GEMET dataset. Section 3 describes how EARTh has been published in the LOD cloud. Section 4 describes the dataset applications. Future steps and conclusions are drawn in Section 5.

2. EARTh Thesaurus

EARTh is a project run by CNR-IIA-EKOLab since 2001. It aims at creating a new thesaurus for the environment extending the GEMET content revising GEMET categorical and thematic structure.

2.1. From GEMET to EARTh

Originally GEMET, developed by CNR-IIA-EKOLab and by German Federal Environmental Agency within an international consortium, was intended to be used as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA). The basic idea for the development of GEMET was to use the best of the presently available excellent multilingual thesauri, in order to save time, energy and funds. GEMET was conceived as a "general" thesaurus, aimed to define a common general language, a core of general terminology for the environment. Specific thesauri and descriptor systems (e.g. on Nature Conservation, on Wastes, on Energy, etc.) have been excluded from the first step of development of the thesaurus and have been taken into account only for their structure and upper level terminology.

Since 2001, CNR-IIA-EKOLab has been performing an overall checking of GEMET in order to improve both its content and its structure. In particular the following activities have been undertaken:

- Quality assessment of GEMET structure and content towards ISO standards on mono- and multilingual thesauri. During GEMET development it was mandatory to take into account some decisions adopted by the consortium and this caused some divergences from standards;
- Assessment of English concept representation vs. source language(s);
- Deletion of incorrect terms and removal of about 1000 terms potentially useful for specific lists, such as name of plants, animals, minerals, etc.;
- Updating the content with new terms and extension of the system of non-descriptors;
- Management of the correspondence of the new terms in British and American English;
- Revision of the thematic structure and development of a new categorical/hierarchical setup to emphasize the different functions of hierarchy in comparison to themes;
- Extension of the horizontal and the vertical relations system;
- Representation of the accessory elements: singular and plural forms; alternate terms, etc.

Besides from GEMET, EARTh terminological content is derived from various mono and multi-lingual sources of controlled environmental terminologies such as UN Environment and Development (1992), Italian Thesaurus of Earth Sciences (2000), Inland Water terminology (2001), Emergency Management Terms Thesaurus (1998/2003) and other terminology collected from reference documents in specific fields or coming from the daily research activity. EARTh currently contains more than 15.000 terms in English and Italian.

2.2. EARTh semantic model

EARTh is based on a multidimensional classificatory and semantic model [6].

The "vertical structure" of the thesaurus is built through a deductive (top-down)–inductive (bottom-up) approach. It is basically mono-hierarchical. It is developed according to a tree semantic model and is based on a system of categories. The first level of categories corresponds to entities, attributes, dynamic

[4] http://www.eionet.europa.eu/gemet

aspects, and dimensions. The vertical structure analyses the primary meaning of the terms and places them in the classificatory-hierarchical tree aiming to orientate the users towards the most "essential" characteristics of terms' semantics.

A thematic organization of terms is elaborated. A theme or a subject is here conceived as a sector of interest that reassembles the terms linked to it. The system of themes as it was conceived is developed according to the specific needs of the applicative context like the classification of terms for the management of environmental information.

"Traditional" thesauri typically provide a poorly differentiated set of relationships between terms, distinguishing only among hierarchical relationships, associative relationships and equivalence relationships. In the EARTh project the standard relationships are being arranged into richer subtypes, whose semantic content is specified. This work is particularly useful dealing with the associative relations (RTs). Typically RTs include a heterogeneous and undifferentiated set of relations, expressing many kinds of association between terms that are not hierarchically based. In EARTh RTs are differentiated into subtypes, thus strengthening the transversal relational structure.

The enrichment of thesaurus relationships and the increased semantic clarification of the relations could enable a better semantic description of web resources and guide users in meaningful information discovery on the web [7].

3. Publishing EARTh as linked data

From the technological point of view, D2R Server[5] has been adopted to map EARTh into the SKOS RDF vocabulary and to make EARTh available as Linked Data. This technological choice has been made considering the following requirements: (i) limited efforts and economical resources were allocated to EARTh publication as Linked Data; (ii) the employment of technologies exhibiting a very low competency barriers was recommendable to show how easily Linked Data can be deployed to environmental community; (iii) EARTh was and is till managed and updated via SuperThes [8] tool which relies on a relational database to store the thesaurus content.

D2R perfectly matches the above requirements. It allows to deploy data from relational databases as SPARQL end points and HTTP dereferenceable linked datasets. It is open sources and free, and users can deploy D2R servers just relying on basic linked data concepts and D2RQ mapping language.
Unfortunately, the original database model adopted by SuperThes is quite complex and stores management information which are not meant to be published. Its direct usage would have turned out in very poor performances and additional efforts defining the D2RQ mapping to SKOS. Next subsections describe the LOD patterns and the database schema adopted exposing EARTh in the Linked Open Data Cloud.

3.1. Adopted Linked Data patterns

In order to make EARTh available as linked data, the following Linked Data patterns [9] has been adopted:

- Naturals Keys for identifier: the internal identifiers for EARTh concepts are adopted as Natural Keys in order to keep a coherence with EARTH's previous (not linked data) releases and usages.
- Label Everything: every concept has its English human-readable name expressed as rdfs:label. So human-readable names can be exploited debugging queries and exploring EARTh.
- Preferred Label: every concept has its preferred label expressed as skos:prefLabel. Both English and Italian lexical representations are provided.
- Materialize Inferences: some of RDF and SKOS entailments have been materialized to support clients with limited processing power. For example, rdfs:labels are obtained as a materialized inferences of English skos:prefLabel. Further materializations can be deployed as discussed in section 3.2.
- Equivalence Links: more than 4000 skos:exactMatch are provided to indicate equivalent URIs between EARTh and GEMET. That has been possible because EARTh is a significant extension of GEMET [10] and explicit references to the GEMET ID were maintained for the concepts shared with GEMET. These equivalence links enable a combined exploitation of EARTh and GEMET taking advantage, at least at some extent, of their respective strengths and complementarities.

3.2. EARTh database schema

The database schema adopted to represent the SKOS elements strongly impacts on the complexity

[5] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server

of the eventual D2RQ mapping specification. It also determines what RDF entailments and SKOS constraints can be ensured by exploiting mechanisms which are native in relational database (e.g., primary and foreign keys, integrity constraints). The database schema illustrated in Fig 1 provides a solution for modelling the database schema in MySQL.

The concept scheme and information (e.g. scheme title, description, thesaurus authors, publishers and top concepts) are represented in the tables "SkosConceptScheme", "SkosConceptSchemeInfo" and "hasTopConcept". Each SKOS concept is associated to a skos:ConceptScheme, thus skos:con-cept(s) are identified by a schema ID ("SkosConcept.skosScheme_ID") and an intrascheme ID ("SkosConcept.ID") in the table "SkosConcept". These identifiers are foreign keys for tables "RDFLabel", "SkosNote" and "SkosSemanticRelation".

Moreover: (i) lexical representations of "skos:concept" such as skos:prefLabel, skos:altLabel can be stored in the table "RDFLabel" specifying distinct "RDFLabel.lexRelType"; (ii) skos:description and skos:notes can be stored in the table "SkosNote" specifying distinct "skosNote.anno-tationType"; (iii) semantic relations (e.g. skos:exactMatch, skos:broadMatch, skos:broader, skos:narrower, skos:related) can be stored in the table "SkosSemanticRelation" specifying distinct "SkosSemanticRelation.relType". Such a modelling solution is designed to explicitly support the subproperty entailments specified in the SKOS recommendation [5]. For example, the statement S22 in the SKOS recommendation defines skos:broader as a subproperty of skos:broaderTransitive and S21 defines skos:broaderTransitive as a subproperty of skos:semanticRelation. That means that each skos:broader is also a skos:semanticRelation. Also skos:narrower, skos:related are defined as subproperties of skos:semanticRelation.

Entailments pertaining to the skos:semanticRelation subproperties can be materialized by specifying in the D2RQ mapping that all the tuples in table "SkosSemanticRelation" have be mapped as skos:semanticRelation values independently from their "relType". Entailments pertaining to skos:broaderTransitive can be materialized by specifying that all the tuples in table "SkosSemanticRelation" having type "broader" must be mapped as skos:broaderTransitive values.

This schema allows to store an unlimited number of languages. Other entailments can be addressed in the storage procedure, which imports data from original sources. For example, for every row inducing a skos:broader relation we could add a row that materializes its inverse skos:narrower. Incidentally, database triggers can be investigated to support entailments when thesauri are dynamically updated.

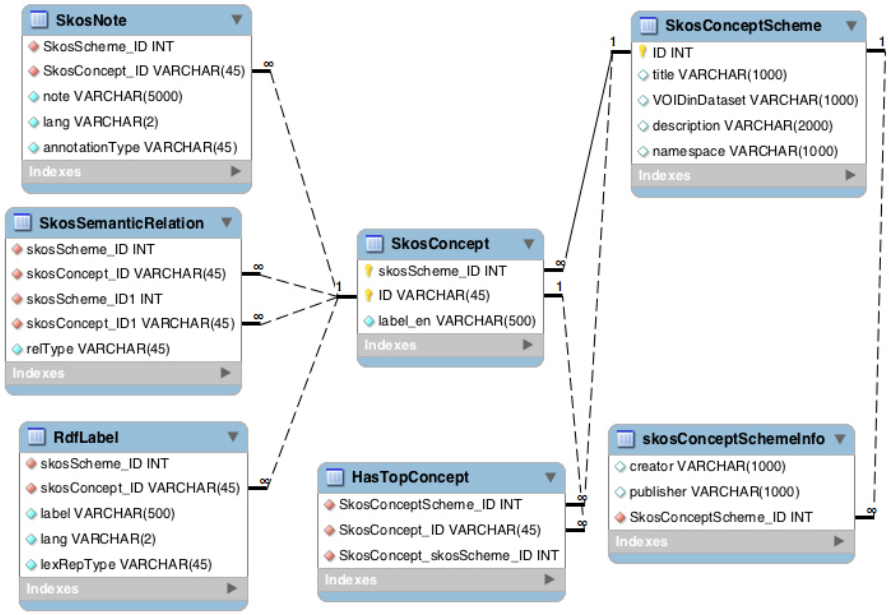

Fig. 1: Database schema adopted to represent EARTh in MySQL.

3.3. EARTh in the LOD cloud

In the late 2011, EARTh has been included in the Linked Open Data (LOD) Cloud.

EARTh content is accessible through (i) HTTP dereferenceable URIs[6]; (ii) RDF/XML dump[7]; (iii) SPARQL end point[8]. EARTh is part of a framework that includes other SKOS linked datasets [11], so accessing EARTh concepts via SPARQL end point requires to make a good use of either (i) the SKOS scheme[9] under which EARTh concepts are grouped, or (ii) the specific URI pattern[10] the EARTh concepts follow. Two examples of simple SPARQL making use of skos:scheme and URI pattern are shown below:

Example1: SPARQL making use of skos:scheme

```
PREFIX skos:
<http://www.w3.org/2004/02/skos/core#>
PREFIX skosScheme:
<http://linkeddata.ge.imati.cnr.it:2020/resource/Skos
ConceptScheme/>

SELECT DISTINCT * WHERE {
  ?s skos:prefLabel ?o.
  ?s skos:inScheme skosScheme:1
} LIMIT 100
```

Example 2: URI pattern

```
PREFIX skos:
<http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT * WHERE {
?s skos:prefLabel ?o.
FILTER
(regex(STR(?s),
"http://linkeddata.ge.imati.cnr.it:2020/resource/EAR
Th/"))} LIMIT 100
```

Statistics pertaining to the number of SKOS concepts and the availability of properties for those concepts are provided in The first column of the table provides information about the number of skos:concept and their SKOS lexical representations. For example, the first row indicates that 14351 skos:concept are available, the second row shows that 14350 of them have a skos:prefLabel in English and 14002 in Italian. The second column shows the number of concepts having at least one occurrence of the indicated SKOS relations. For example, the first row indicates that 14350 concepts have a skos:inScheme property.

Table 1. The first column of the table provides information about the number of skos:concept and their SKOS lexical representations. For example, the first row indicates that 14351 skos:concept are available, the second row shows that 14350 of them have a skos:prefLabel in English and 14002 in Italian. The second column shows the number of concepts having at least one occurrence of the indicated SKOS relations. For example, the first row indicates that 14350 concepts have a skos:inScheme property.

Table 1

Property	#	Property	#
skos:concept	14351	**skos:inScheme**	14350
skos:prefLabel	14350 **(en)** 14002 **(it)**	**skos:broader or skos:narrower**	11664
rdfs:label	14350 **(en)**	**skos:exactMatch to GEMET**	4365
skos:definition	6332 **(en)** 5883 **(it)**	**skos:related**	4083
skos:altLabel	1198 **(en)** 853 **(it)**		

Further details pertaining to linkset and accessibility are available in VOID description[11] and dataHub[12].

EARTh is available under by-nc-nd creative commons licence[13], which grants the right to copy, distribute and transmit it for non-commercial purposes, but implies explicit attribution of work and forbids derived works.

4. EARTh application/relevance

EARTh represents an environmental thesaurus to be utilized for different purposes being able to combine the search of stable logical and conceptual basis with a flexibility towards different applications, to ensure an optimal conceptual coverage, to be aware of the cultural dimension of knowledge organization, to allow different levels of comprehensibility and applicability for users with different expertise.

[6] http://linkeddata.ge.imati.cnr.it:2020/directory/EARTh
[7] http://purl.oclc.org/net/DumpEarthRDF
[8] http://linkeddata.ge.imati.cnr.it:2020/sparql
[9] http://linkeddata.ge.imati.cnr.it:2020/resource/SkosConceptScheme/1
[10] http://linkeddata.ge.imati.cnr.it:2020/resource/EARTh/.*
[11] http://purl.org/NET/Earth.ttl
[12] http://thedatahub.org/dataset/environmental-applications-reference-thesaurus
[13] http://creativecommons.org/licenses/by-nc-nd/3.0/

EARTh relevance is recognised by the Italian Ministry for the Environment as a glossary for the management of information on waste management[14] and by the Italian Environmental Agency[15] for the indexing technical and scientific documents.

Its availability as linked dataset has raised its relevance at international level, especially in European projects (e.g., NatureSDIplus[16], NESIS[17], EGIDA[18]). In particular in NatureSDIplus, EARTh has been recognised as the pivotal general-purpose thesaurus for the design of a Common Thesaurus Framework for Nature Conservation integrating different thesauri for specific data themes [11].

5. Conclusion and Future work

The paper illustrates the main characteristics of the linked dataset EARTh, an environmental thesaurus extending GEMET that promises to become pivotal for data sharing in the environmental domain. EARTh content continuously evolves as result of CNR-IIA-EKOLab's research activity, whist EARTh Linked Data releases are provided once a year by CNR-IMATI.

Future activity will include improvement in terms of EARTh content as well as its Linked Data publication. Concerning the content, an overall revision of the thesaurus structure and content is currently undergoing as consequence the recent publication of ISO 25964-1:2011 [12]. The number of concepts and lexical correspondents provided by EARTh are expected to increase also as a consequence of a larger adoption by environmental communities. Concerning the Linked Data publication, novel releases are expected to overcome current limitations. In particular, (i) some of the materializations (e.g., entailments of skos:semanticRelation) described in this paper are not yet deployed and will be made available in the future releases; (ii) RT properties, which have been indistinctly mapped into skos:related in order to avoid the adoption of user-defined RDF vocabularies, will be differentiated as in the original version of EARTh; (iii) EARTh connection with other environment-related thesauri will be strengthened. In particular concerning the concepts that are shared with GEMET, the connection activity will start analysing the GEMET's linkset to other vocabularies. Eventually, new linkset will be created by combining the CNR-IIA-EKOLab's domain expertise with the usage of link discovery tools like SILK.

[14] http://www.osservatorionazionalerifiuti.it
[15] http://www.envidocnet.isprambiente.it/INDEKS/public/welcome.do
[16] http://www.nature-sdi.eu/
[17] http://www.nesis.eu/

Acknowledgements

This activity has been partially supported within the EU project NatureSDIplus (ECP-2007-GEO-317007), eContentplus program.

References

[1]Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007. http://inspire.jrc.ec.europa.eu/

[2]De Groof, H., SEIS, Status of Play, http://www.nesis.eu/proceedings/krakow/degroof.pdf, 2010,

[3] Bandholtz T., Fock J., Legat R., Nagy M., Schleidt K., Plini P., 2009. Shared Terminology for the Shared Environmental Information System. Environmental Informatics and Industrial Environmental Protection: Concepts, Methods and Tools, 23rd International Conference on Informatics for Environmental Protection., Volume 1. Shaker, Aachen, pp. 123-127

[4] T. Berners-Lee, "Design Issues: Linked Data," 2006, http://www.w3.org/DesignIssues/LinkedData.html,.

[5] SKOS (2009). Simple Knowledge Organization System Reference, W3C, Recommendation, at http://www.w3.org/TR/skos-reference.

[6] Mazzocchi F., De Santis B., Tiberi M., Plini, P., 2007. Relational Semantics in thesauri: Some Remarks at Theoretical and Practical Levels. Knowledge Organization, vol. 34 (4). pp. 197-214.

[7] Soergel D., Lauser B., Liang A., Fisseha F., Keizer J. and Katz S.. Reengineering thesauri for new applications: the AGROVOC Example. Journal of digital information 4 issue. 4. Article No. 257, (2004)

[8] Felluga, B., Batschi W.D. (Eds.), 1999. GEMET, General European Multilingual Environmental Thesaurus. Version 2.0. European Environmental Agency, Copenhagen.

[9] Dodds, L, Davis I, Linked Data Patterns. A pattern catalogue for modelling, publishing , and consuming Linked Data. Retrieved from http://patterns.dataincubator.org/book/

[10] Plini, P., Franco, S. D., & Santis, V. D. (2009). A state-of-the-art of Italian National Research Council (CNR) activities in the area of terminology and thesauri. In J. Hřebíček, J. Hradec, E. Pelikán, O. Mírovský, W. Pillmann, I. Holoubek, & T. Bandholtz (Eds.), European conference of the Czech Presidency of the Council of the EU TOWARDS eENVIRONMENT Opportunities of SEIS and SISE: Integrating Environmental Knowledge in Europe

[11] De Martino M., Albertoni R.; A multilingual/multicultural semantic-based approach to improve Data Sharing in a SDI for Nature Conservation, International Journal of Spatial Data Infrastructures Research, Vol.6, ISSN 1725-0463, pp. 206-233, (2011)

[12] ISO 25964-1 Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval.