The original publication is available http://www.springerlink.com/.
Riccardo Albertoni, Monica De Martino, Paola Podestà
Environmental Thesauri under the Lens of Reusability,
EGOVIS 2014: Electronic Government and the Information Systems Perspective pp 222-236

Volume 8650 of the book series Lecture Notes in Computer Science (LNCS)

Kő A., Francesconi E. (eds)
2014,
ISBN: 978-3-319-10177-4
Doi: 10.1007/978-3-319-10178-1_18

Environmental thesauri under the lens of reusability

Riccardo Albertoni, Monica De Martino, and Paola Podestà

Istituto di Matematica Applicata e Tecnologie Informatiche
Consiglio Nazionale delle Ricerche,
Via De Marini, 6, 16149 Genova, Italy
{albertoni,demartino,podesta}@ge.imati.cnr.it

Abstract. The development of a Spatial Data Infrastructure (SDI) at European level is strategic to answer the needs of environmental management requested by the European, national and local policies. Several European projects and initiatives aim to share, integrate and make accessible large amount of environmental data in order to overcome cross-border/language/cultural barriers. To this purpose, environmental thesauri are used as shared nomenclatures in metadata compilation and information discovery, and they are increasingly made available on the web. This paper provides a methodological approach for creating a catalogue of the environmental thesauri available on the web and assessing their reusability with respect to domain independent criteria. It highlights critical issues providing some recommendations for improving thesauri reusability.

Keywords: Environmental thesauri, Linked Data, Spatial Data Infrastructure, Open Government, metadata.

1 Introduction

In recent years, 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, nevertheless, an effective sharing of these resources is still an open issue due to the intrinsic multicultural and multilingual nature of the environmental domain. Thus, the development of a Spatial Data Infrastructure (SDI) at European level requires to deploy geographic data in a standardized way and with common nomenclatures. Different communities having a large spectrum of competencies are involved in the treatment and the management of geographical information, consequently SDI deals with several thesauri in order to deeply cover such a variety of competencies. Currently several thesauri for the Environment are shared

[1] http://inspire.ec.europa.eu/
[2] http://ec.europa.eu/environment/seis/

in the web embodying different points of view and different ways of conceptualization. These thesauri are precious and their exploitation within a SDI for metadata compilation and data discovery is critical.

Our experience in the management of Environment-related thesauri started in the European project NatureSDIplus[3] aimed at supporting the implementation of INSPIRE. The goal of this project has been the harmonization and the integration, at European level, of the datasets on nature conservation, available on the web, to better exploit and access them. This has been a challenging task due to the several existing Knowledge Organization Systems (KOS), such as taxonomies, thesauri, code lists, gazetteers, etc... Moreover, the development of new resources might result in a huge waste of effort attempting to reinvent the wheel, and in a possible increasing of the information redundancy. Thus, the approach in the NatureSDIplus has been the creation of a framework for the integration of existing KOS, using Linked Data best practices, in order to harmonize the data (and metadata) entry and to support the information retrieval using metadata in a Metadata Information System. Following the agreement and the interest for the thesaurus framework shown inside the EU Community, a further activity, in the new ongoing EU project eENVplus[4], has been planned to enrich the thesaurus framework adding further environmental thesauri in order to improve the existing services to overcome cross-border and language barriers.

In recent years, several organizations have provided their KOS on the web using the Simple Knowledge Organization Systems (SKOS) [11] as common data model and they have published some of these SKOS as Linked Data. Considering the perspective of the integration of existing KOS in an SDI, the activity concerning the identification of the reusable KOS, is critical.

Some recent papers also contributes in addressing the reusability of environmental thesauri of considering different points of view. The paper [10] presents a survey for understanding the modelling style in terms of shape, size and depth of the vocabulary structured as SKOS vocabulary on the web. It mainly focuses on the usage of SKOS constructs, SKOS semantic relations and lexical labels. In [15] a framework for the automated assessment and correction of common potential quality issues in SKOS vocabularies is proposed. The quality measures defined in the framework consider not only structural issues, but also labelling and documentation issues such as missing or overlapping labels, and also Linked Data specific issues, such as broken links, missing inlinks, invalid URIs. Instead, [12] presents an analysis of the KOS available on the web which is independent from their SKOS structures. The considered KOS are identified using journal and scientific sources. Then, they are classified considering the type (thesaurus, ontology and glossary), the covered science domain, the continent of origins and the date on which they are made available on-line.

In this paper we present an approach to identify a set of environmental thesauri available on the web and to assess their reusability, in terms of easiness to access and to exploit their content. To this purpose, first of all, we consider

[3] http://www.nature-sdi.eu/
[4] http://www.eenvplus.eu

the best practices for publishing Linked Data (see [7]), based on the 5 Star Linked Data principles (5 star LD [2]), that sets out a series of best practices designed to facilitate development and delivery of government data as Linked Data. Moreover we refer also to the papers [5, 14] stating respectively that the adoption of Linked Data best practices jointly with SKOS and the type of licence are essential in the deployment of a resource in the web. Thus, we address the assessment of reusability considering the openness of licence and the compliance to the 5 star LD, stressing on, for the latter, the deployment of *derefererceable* HTTP URIs as identifiers for resources. Licence and HTTP *dereferenceability* are central prerequisites for every scenario of reuse and they are crucial for interlinking among structured data, but they are not considered at all in [10], [12] and [15].

The contributions of the papers are the following:

- the definition of a methodological approach which includes the employment of different investigation strategies to collect a set of possibly well known terminological resources for the Environment among those available in the web;
- the synthesis and explication of a set of reusability criteria which, although quite settled in the Linked Data community, are not yet fully received by environmental thesauri producers and publisher;
- a "reference" catalogue of thesauri which can be exploited by data users and applications in the Environment domain;
- the reusability assessment of the thesauri in the catalogue and the discussion of issues arising from the reusability analysis and some recommendations, which might result interesting for thesauri users and publishers for screening the thesauri they want to adopt or for improving the reusability of their own thesauri.

2 Introduction to the methodological approach

This section outlines the main steps and the characteristics of the methodology adopted aimed at identifying the environmental thesauri to be evaluated in the reusability perspective. The methodology is defined by a multi-task process as represented in the workflow in Fig. 1. It is characterized by three main phases:

- *Phase I. Resource identification and cataloguing:* identification of the available thesauri for the Environment and creation of the thesaurus catalogue.
- *Phase II. Identification of reusability criteria:* identification and formalization of technological criteria able to evaluate the reusability.
- *Phase III. Evaluation of thesauri:* assessment of the reusability of the thesauri according to the criteria previously identified.

It is important to highlight that different communities connected with the environmental domain have been involved in a continuous interaction in order to set up the initial set of thesauri and to sort out doubts and issues arising during all the three phases of the process. In the following we describe the three phases in detail.

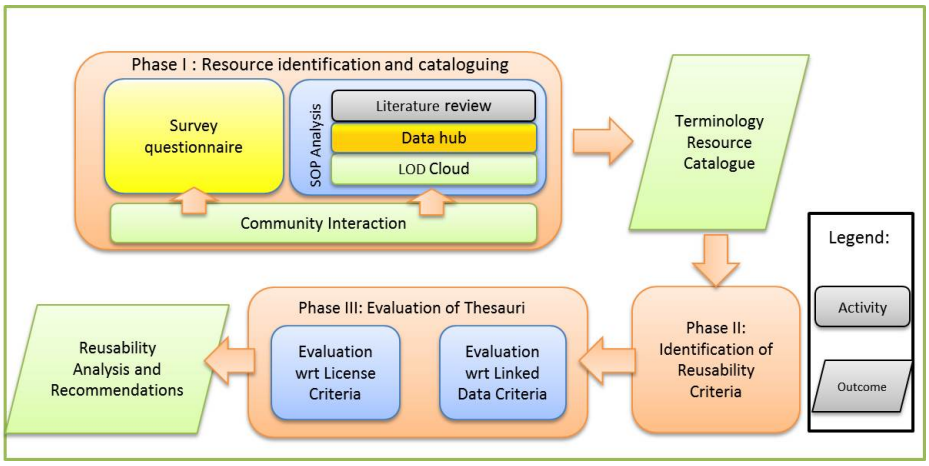

Fig. 1. Workflow of the methodological approach.

3 Resource identification and cataloguing

This activity aims at identifying and collecting as many different environmental thesauri as possible in order to perform on them some representative analysis concerning reusability. The catalogue does not want to be exhaustive of all existing terminologies of the environmental domain. However the multi-strategy process adopted for identifying the thesauri entails the catalogue as a good "reference" catalogue, representative of the well- and quite-known environmental thesauri available on the web and possibly in Linked Data.

3.1 Resource identification

In order to identify the available terminologies a multi-strategy process of investigations has been adopted, considering and combining different types of information sources. The strategies adopted are: (i) an on-line questionnaire, (ii) a State of Play analysis (SoP), and, (iii) the direct interaction with different environmental communities.

On-line questionnaire. An on-line questionnaire entitled *Thesaurus survey* has been created in order to identify a preliminary set of terminologies. It has been distributed among several environmental communities, such as National and European environmental agencies and terminological experts in the community of Networked Knowledge Organization (NKOS). The questionnaire has totally 85 questions divided in five sections. The information requested can be summarized into three main groups:

- general evaluation of the user′s skills in thesauri;
- identification of new terminological resources;

– collection of technical details about the suggested terminological resources (e.g., licence, available format).

In order to give a weight to the information suggested in the questionnaire, we have evaluated also the users skills and experience in using thesauri. The total number of responses has been 54 and about the 70% (37 units) of the responses are from users with experience about thesaurus, this guarantee a good reliability of the questionnaire suggestions.

State of Play analysis (SoP). The state of play aims at identifying the available terminologies that may be accessed through the web in order to complement the answers provided by the questionnaire. The methodology adopted is based on an Internet survey conducted using well-known search engines/platforms, the scientific literature and the interaction with the community. In particular:

– *Scientific literature.* This category includes mainly papers published in scientific international journals or conference proceedings relevant in the fields of Linked Data and Semantic Web. In particular we have focused on the Semantic Web Journal (SWJ), which has recently started a section dedicated to the descriptions of impacting Linked Data Datasets. Terminological resources included in this section of the SWJ are usually of high quality and technically validated by the community of Linked Data. We have also considered a previous survey on environmental terminologies presented in [12].
– *The datahub.* The datahub is a platform developed to share open datasets through a specific section for Linked Data. We have searched in the datahub all the terminolgies associated to the keywords *thesaurus* and *skos*. Among them we have considered those thematically related to environmental domain and also those interlinked to one of the main thesaurus players in the Environment (e.g., GEMET, AGROVOC, EARTh).
– *LOD Cloud.* This category includes terminological resources shared in the datahub and included in the LOD Cloud. The LOD Cloud diagram represents datasets published by the Linking Open Data project from 2007-2011. Terminological resources have been marked as included in the LOD Cloud according to the analysis available in http://validator.lod-cloud.net/.

Community interaction. Different communities related to the environmental sciences have been involved in compiling the on-line questionnaire and in sorting out issues arising during the SoP investigation. In particular, the involved communities are:

– Public and private environmental stakeholders;
– Members of National Environment Agency of several European country as well as the European Environment Agency (EEA) and the Joint Research Centre (JRC - European Commission);
– Terminological thesauri experts from the mailing list Ecoterm and community of experts on Networked Knowledge Organization (NKOS).

The coverage of these multi-strategy process seems to be quite adequate, since it stresses and combines quite all the available type of information sources:

(i) the web; (ii) the literature, focusing in particular on previous survey on environmental thesauri [12] and on the Semantic Web Journal; and (iii) the community, through the on-line questionnaire and the continuous interaction with the environmental domain experts.

3.2 Reference catalogue of thesauri

The multi-strategy investigation has resulted in a collection of different types of terminological resources. In fact, even if our research has been focused only on thesauri, indications returned by the multi-strategy process has also included codelists, ontologies, taxonomic datasets, datasets, gazetteers, schema/rdf vocabularies, glossary, vocabularies for a total of 62 resources. This is probably due to an inappropriate use of the term "thesaurus" among the communities.

In this paper, we decide to consider the thesauri, that is a controlled vocabulary of terms where semantic relations (hierarchical, associative, equivalence) between terms are explicitly declared. The total number of collected thesauri is 24. Table 1 shows the catalogue of thesauri providing: (i) the resource acronym; (ii) the resource description presenting the name of the thesaurus and some descriptive information (URL, datahub ID, scientific reference, licence); (iii) the provenance indicating the sources from which the thesauri has been collected, i.e., the questionnaire (Q), the LOD Cloud (LC), the SWJ dataset section (L), the datahub (DH) and the community suggestions (C).

Let′s note that the adoption of a multi-strategy investigation allows to detect the presence of the same thesaurus in different sources provide a thumb rule of its "popularity" in environmental and Linked Data communities.

Resource acronym	**Resource description**	**Provenance**
ADL FTT	**Alexandria Digital Library Feature Type Thesaurus** **URL:**http://www.alexandria.ucsb.edu/ lhill/FeatureTypes/ver070302/ **Licence:**http://www.alexandria.ucsb.edu/gazetteer/#licensing	C
AGROVOC	**AGROVOC** **URL:** http://aims.fao.org/standards/agrovoc **Bibliographic Reference:** [3] **Datahub ID:** agrovoc-skos **Licence:**http://creativecommons.org/licenses/by/3.0	Q, DH, L, LC, C
EARTh	**Environmental Applications Reference Thesaurus** **URL:** http://linkeddata.ge.imati.cnr.it:2020/ **Bibliographic Reference:** [1] **Datahub ID:** environmental-applications-reference-thesaurus **Licence:**http://creativecommons.org/licenses/by-nc-nd/3.0/	DH, L, LC
EcoLexicon	**EcoLexicon** **URL:** http://ecolexicon.ugr.es/visual/index_ en.html **Bibliographic Reference:** [6] **Licence:** Not found	Q
EnvThes	**EnvThes - Environmental Thesaurus** **URL:** http://vocabs.lter-europe.net/ EnvThes3.html **Licence:**In progress	Q
EOSterm	**Earth Observation Systems Thesaurus** **URL:** http://thesaurusonline.iia.cnr.it/ tematres/eosterm **Reference:** [8] **Licence:**http://creativecommons.org/licenses/by-nc-nd/3.0/	Q
EuroVoc	**EuroVoc Multilingual Thesaurus of the European Union** **URL:**http://eurovoc.europa.eu/ drupal/ **Datahub ID:** eurovoc-in-skos **Licence:**http://eurovoc.europa.eu/drupal/?q=legalnotice&cl=en	DH, C

GEMET	**GEneral Multilingual Environmental Thesaurus**	Q DH, LC
	URL: http://www.eionet.europa.eu/ gemet/	
	Datahub ID: gemet	
	Licence:http://creativecommons.org/licenses/by/2.5/dk/	
GBA	**Geological Survey of Austria (GBA)- thesaurus**	DH, LC
	URL:http://resource.geolba.ac.a	
	Datahub ID:geological-survey-of-austria-thesaurus	
	Licence:http://opendefinition.org/licenses/cc-by-sa	
ICAN	**ICAN demonstrator thesaurus**	C
	URL:http://mmisw.org/ont/ican/thesaurus	
	Licence: Not found	
Inter WATER	**InterWATER Thesaurus**	C
	URL:http://thesaurus.ircwash.net/	
	Licence:http://creativecommons.org/licenses/by-nc-sa/3.0/nl/deed.en	
IUGS-CGI	**IUGS-CGI Multi-Lingual Thesaurus of Geosciences**	C
	URL:http://www.cgi-iugs.org/tech_ collaboration/ thesaurus.html	
	Licence: In progress	
NALT	**The U.S. National Agricultural Library Thesaurus**	Q, DH, LC
	URL: http://agclass.nal.usda.gov/	
	Datahub ID: nalt	
	Licence: http://www.nal.usda.gov/web-policies-and-important-links#NAL%20Agricultural%20Thesaurus%20and%20Glossary	
NERC NVS2.0	**NERC Vocabulary Server version 2.0**	Q, DH
	URL: http://vocab.nerc.ac.uk	
	Datahub ID: nvs	
	Licence:http://www.nationalarchives.gov.uk/doc/open-government-license/version/2/	
SEMIDE	**SEMIDE Thesaurus**	C
	URL: http://www.emwis.net/portal_ thesaurus	
	Licence:http://www.emwis.net/about/copyright_html	
SnowTerm	**SnowTerm**	Q
	URL: http://192.167.230.177/tematres/snowterm/	
	Bibliographic Reference: [13]	
	Licence:http://creativecommons.org/licenses/by-nc-nd/3.0/	
SoilThes	**SoilThes**	Q
	URL: https://secure.umweltbundesamt.at/soil/en/ collections/SoilCore_0.htm	
	Licence: http://creativecommons.org/publicdomain/zero/1.0/	
STW	**STW Thesaurus for Economics**	DH, LC
	URL:http://zbw.eu/stw/versions/latest/	
	Datahub ID: stw-thesaurus-for-economics	
	Licence:http://creativecommons.org/licenses/by-nc/2.0/	
TheSoz	**TheSoz (Thesaurus for the Social Sciences)**	DH; L
	URL: http://lod.gesis.org/thesoz/	
	Bibliographic Reference: [16]	
	Datahub ID: gesis-thesoz	
	Licence:http://creativecommons.org/licenses/by-nc-nd/3.0/de/	
ThIST	**Italian Thesaurus of Sciences of the Earth**	Q
	URL:http://sgi.isprambiente.it/OnThist/servlet/onthist	
	Bibliographic Reference: [4]	
	Licence: In progress	
UMTHES	**UMweltTHESaurus**	DH
	URL:http://data.uba.de/umt/de.html	
	Datahub ID:umthes	
	Licence: http://opendefinition.org/licenses/cc-by/	
UNESCO	**UNESCO Thesaurus**	DH
	URL: http://databases.unesco.org/thesaurus	
	Datahub ID: unescothes	
	Licence:http://creativecommons.org/licenses/by-nc/2.0/	
U.S.G.S.	**United States Geological survey (Science,Themes and Subject)**	C
	URL: http://www.usgs.gov/science/about/	
	Licence: Not found	
WQPB	**WQPB (Water Quality Library Thesaurus)**	C
	URL: http://svc.mt.gov/deq/wqlibrarysearch/Thesaurus.pdf	
	Licence: Not found	

Table 1: Reference Catalogue of 24 Thesauri.

4 Identification of reusability criteria

This section presents the formalization of the criteria adopted for the evaluation of thesaurus reusability. We consider two different criteria, one based on the 5 star LD principles defined by Tim Berners-Lee in [2] and the other based on the type of licence under which the thesaurus is released. They are explained in detail in the following.

4.1 5 star LD principles

In this section we present the formalization of the criteria for assessing the thesaurus compliance with 5 star LD classification (see [2]).

In our analysis special attention is paid to dereferenceability of the URI associated to concepts in the thesaurus. Dereferenceable URIs are the mandatory prerequisite for Linked Data, in fact, without them, it is not possible to check what is attached to the URI, and thus the identifiers are not truly reusable. In particular, the provision of thesaurus concepts without dereferenceable URIs restricts the third-parties possibility (i) to check authoritativeness of information associated to thesaurus concepts; (ii) to exploit mappings among thesauri concepts in order to discover further information in a follow-your-nose fashion. Coherently with the importance of HTTP dereferenceable URI in the Linked Data design issues, we have assigned 4 stars only to thesauri whose identifiers are HTTP dereferenceable and return RDF/XML encoding. Thus, we have detailed the 5 star LD classification proposed in [2] adding the values 3.5 and 3.9 between 3 and 4 stars, as follows:

- *1 star:* resources available on the web (whatever format);
- *2 stars:* resources available as machine-readable structured data (e.g., Excel instead of image scan of a table);
- *3 stars:* as 2 stars plus non-proprietary format (e.g., CSV instead of Excel);
- *3.5 stars:* resources available as RDF dump without dereferenceable HTTP URI;
- *3.9 stars:* resources provided as RDFa (RDF embedded in XHTML) or SPARQL end point which are very close to be Linked Data ready but still without dereferenceable HTTP URI.
- *4 stars:* all the above plus, use open standards from W3C (RDF and SPARQL) and HTTP dereferenceable URI to identify things, so that people can point at published resources;
- *5 stars:* all the above, plus links to other people′s data to provide context.

In order to correctly evaluate the HTTP dereferenceability, concept URIs have been tested following the standard procedure detailed in the second section of Heath′s book [9]. This procedure relies on the basics of the HTTP protocol: it sends a HTTP GET request for the URI indicating RDF/XML as preferred representation, and then it interprets the server response following any `303 redirects` till a `200 OK` is reached. If the `200 OK` is reached and a RDF returned then the URI is considered HTTP dereferenceable. Otherwise, it isn't.

4.2 Licence criteria

This section presents the licence criteria considering the categories presented in [14] that are based on some existing and well-known type of licences, such as the framework defined by Creative Commons. We decide to consider this framework since it provides an exhaustive coverage, the licences are identifiable by URIs and they are intended for general intellectual works. In the following we explain the formalization presented in Table 2.

- **Licence (acronym)/Characteristics.** We have slightly changed the categories defined in [14]. In fact we have divided the category *Not specified* distinguishing the subcases *Not found* and *In progress* in order to capture all the cases we have faced during the search of licence information. The category considered in the evaluation are detailed in the following.
  - *Public Domain Licences (CC0).* They waive all the possible intellectual property and neighboring rights of the resources.
  - *Attribution Licences (CC-BY).* They waive all the possible rights, requiring only the mere attribution.
  - *Share-alike Licences (CC-SA).* The rights are also waived requiring that derived or adapted resources keep the same licence.
  - *With restrictions (CC-NC, CC-ND, CC-NC-ND).* These licences present some restrictions in particular: (i) non-commercial (NC) means that the exploitation of a resource and its derived work must be non-commercial; (ii) non derivative (ND) allows for redistribution, commercial and non-commercial exploitation, as long as it is passed along unchanged and in whole, with credit to creators/right-holders.
  - *In progress (Pr).* In this case, there is an explicit indication on the web site that the licence is under construction or we have a direct knowledge that thesaurus licence is going to be defined soon. *In progress* is a quite common situation: often a thesaurus is a result of the integration of work of different actors, thus it is not easy to choose a licence model which fits for all the contributors.
  - *Not found (NF).* No licence has been found in the website or elsewhere.
- **Licence reusability evaluation.** We have assigned to each type of licence a value meaning the level of reusability of the resource allowed by the licence (1=low reusability, 5= high reusability). As shown in the Table 2 the most important categories are those referring to open licences without severe restrictions (CC0, CC-BY, CC-SA), since they allow the complete reuse, transformation and the publication of a resource.

5 Evaluation of reusability

The thesauri collected in the reference catalogue have been analysed and evaluated with respect to the reusability criteria. In the following we present the evaluation of the thesauri considering the 5 star LD priciples, the licence criteria and the overall results of the analysis highlighting critical issues.

Licence (acronym)	Characteristics	Licence reusability evaluation
Public Domain (CC0)	All the rights have been waived	5
Attribution (CC-BY)	Attribution is required	4.5
Share alike (CC-SA)	Copyleft licence	4
With restrictions (CC-NC , CC-ND, CC-NC-ND)	More severe restrictions	3.5
Closed (CR)	Closed licence	3
In progress (Pr)	Licence is going to be defined soon	2
Not found (NF)	No licence has been found in the website	1

Table 2. Definition of the adopted categories of licence and the levels of reusability of the resource allowed by the licence.

5.1 Evaluation wrt 5 star LD priciples

The evaluation of the thesaurus compliance with respect to the 5 star LD principles is presented in Table 3. The following groups of thesauri can be outlined:

- *Linked Data ready thesauri (LD ready).* This group contains thesauri published according to the Linked Data best practices and exposing dereferenceable concept URIs returning the proper RDF/XML fragments (i.e., LD stars >=4).
- *RDF ready thesauri (RDF ready).* It considers thesauri for which some sort of RDF document is provided but without exposing HTTP dereferenceable URI for their concepts (i.e., 3< LD stars <4).
- *Other format thesauri (Other).* It includes thesauri made available in other format than RDF (i.e., LD stars<=3).

Moreover, about 45% of the considered thesauri (11 out of 24) falls in the first category *Linked Data ready thesauri*. In particular, we find that all the thesauri in this category deploy SKOS as RDF vocabulary. Some of them deploy ad hoc RDF vocabularies or ontologies together with SKOS, for example AGROVOC exploits AGRONTOLOGY, an ontology that basically extends `skos:related` properties with domain dependent relations such as `afflicts /affect`, `controls /isControlledBy`. Six thesauri in this category are already interlinked with third parties thesauri (i.e., LD stars >=5). Then, about the 33% of the thesauri (8 out of 24) falls in the second category. These thesauri already provide some sort of RDF document for their concepts so their exposition as Linked Data is probably under consideration or in progress. All the thesauri in the second category, but ADL FTT, deploy SKOS as RDF vocabulary. ADL FTT deploys an experimental RDF version that is dated back to 2002 and is based on undocumented ESRI vocabulary, probably one of the first attempts to define a RDF vocabulary for thesauri which has been eventually superseded by SKOS. ThIST, EOSTerm, and SnowTerm are classified as 3.5 stars because already available as SKOS-RDF but without HTTP dereferenceability. Moreover, ThIST, EOSTerm, and SnowTerm do not provide a complete SKOS/RDF dump

5 star evaluation	Thesaurus acronym
5	SoilThes, GEMET, AGROVOC, NERC NVS2.0 ,GBA, TheSoz, EARTh, EnvThes
4	NALT, UNESCO, ICAN
3.9	STW
3.5	EuroVoc, UMTHES, SnowTerm, EOSterm, ThIST, ADL FTT, U.S.G.S.
2	IUGS-CGI
1	SEMIDE, InterWATER, EcoLexicon, WQPB

Table 3. Analysis of the thesauri in the catalogue according with 5 star LD principles.

of their overall set of concepts. They provide only a RDF fragment for each concept which is downloadable from HTML concept page or via in-house web application. Similarly, UMTHES provides RDF fragments accessible from the HTML concept page, but it also implements HTTP 303 redirection to adhere to the Linked Data best practices. Unfortunately, when UMTHES concept URIs are dereferenced asking for RDF/XML document, the URIs redirect to HTML pages and not to the proper RDF fragments. Another interesting example is STW Thesaurus for Economics evaluated with 3.9 stars since its set of concepts is complete available as RDFa but without any HTTP dereferenceable concept URIs. Finally, there is the group of thesauri that are not yet available as RDF (5 out of 24). In this group we can distinguish between thesauri accessible on a machine-readable format such as IUG-CGI Thes. of Geoscience, that is available as Excel, and thesauri like SEMIDE, EcoLexicon, IUGS-CGI which are available only embedded in a web portal or as PDF.

5.2 Evaluation wrt Licence criteria

The licence evaluation requires first of all a careful analysis of each thesaurus licence in order to match it with the main characteristics of the Creative Common categories explained in Table 2.

In Table 4 the sign X in a column implies that the licence of thesaurus has such specific characteristic. Beside X, in parentheses, we provide further details:

- (1.0)/(2.5)/(2.0)/(3.0): it is the number of the version of the licence;
- (dh): it indicates that the URL of the licence has been found on datahub platform. For example for the thesauri GBA and UMTHES the following situations arise:
  - the URL points to an HTML pages with links to different versions of the same licence (e.g., http://opendefinition.org/licenses/cc-by-sa/). Thus, it is not possible to identify the correct version (GBA, UMTHES);

Licence evaluation	Thesaurus acronym	CC-BY	CC-NC	CC-SA	CC-ND	CC0	CR	NF	Pr
5	SoilThes					X(1.0)			
4.5	GEMET	X (2.5)							
	AGROVOC	X(3.0)							
	NERC NVS2.0	X(nstd)							
	NALT	X(nstd)							
	EuroVoc	X							
	UMTHES	X(dh)							
4	GBA	X(dh)		X(dh)					
3.5	TheSoz	X(3.0)	X(3.0)		X(3.0)				
	EARTh	X(3.0)	X(3.0)		X(3.0)				
	UNESCO	X(2.0)	X(2.0)						
	EOSterm	X (2.5)	X (2.5)		X (2.5)				
	SnowTerm	X(3.0)	X(3.0)		X(3.0)				
	SEMIDE	X(nstd)			X(nstd)				
	Inter WATER	X (3.0)	X (3.0)	X (3.0)					
	STW	X(2.0)	X(2.0)						
2	EnvThes								X
	IUGS-CGI								X
	ADL FTT								X
	ThIST								X
1	ICAN							X	
	U.S.G.S.							X	
	EcoLexicon							X	
	WQPB							X	

Table 4. Licence analysis of thesauri in the reference catalogue.

  - on the official website of the thesaurus no licence is found. In this case we are not sure that the licence on datahub is correct, since in the past the datahub was a collaborative platform where everyone could modify the information associated to the shared resources;

– (nstd): it indicates that the licence does not refer to a standard framework, thus, it may be difficult to identify all the characteristics of the licence itself. In particular, for the SEMIDE thesaurus the sentence "Reproduction is authorized, provided the source is acknowledged, except where otherwise stated" is ambiguous since it is no immediately clear if derivative works (remix, transformation ect) are authorized. On the other side, for NALT and NERC NVS2.0 it is more simple to categorize the main characteristics, even if it necessary a careful examination of the licences content.

Among the thesauri included in the category *In progress*, we distinguish two cases. In one case the legal notice on the website of the considered resource declares explicitly that the licence is under definition (ADL FTT). In the other case we know that the licence will be defined soon because we are in contact with the developers of the thesaurus (e.g., for EnvThes, ThIST, IUGS-CGI). Then, we have assigned to each thesaurus licence a reusability value according with Table 2. Notice that, if the thesaurus licence matches more than one characteristics we have considered the minimum of the different reusability values associated to the considered characteristics. For example, the licence of the thesaurus *TheSoz*

	LD ready	RDF ready	Other
Open Licenced	SoilThes, GEMET, AGROVOC, NERC NVS2.0, GBA, NALT	EuroVoc, UMTHES	
Partially Open Licenced	TheSoz, EARTh, UNESCO	STW, SnowTerm, EOSterm	SEMIDE, InterWATER
Closed Licenced	EnvThes, ICAN	ThIST, U.S.G.S., ADL FTT	IUGS-CGI , EcoLexicon, WQPB

Table 5. Analysis of the thesauri with the macro-categories identified for LD stars and licence.

includes the clauses CC-BY (its reusability value is 4.5), CC-ND and CC-NC (their reusability value is 3.5 for both), thus we assign to *TheSoz* the value 3.5.

Using the information in Table 4, we can group the thesauri in three categories:

- *Open Licenced Thesauri*. It includes highly reusable thesauri that are released under public domain, attribution or share-alike licences. They can be modified and extended as needed and deployed in commercial and non-commercial context (licence evaluation>=4).
- *Partially Open Licenced Thesauri*. This group contains thesauri licenced with some further restrictions in reusability (licence evaluation=3.5).
- *Closed Licenced Thesauri*. It considers thesauri in which licence forbids the free reuse or for which a licence is not provided yet (licence evaluation<3.5).

The thesauri in the catalogue are equally distributed among these three categories, that means that only the 33% of thesauri considered are truly open licenced. Within the Partially Open Licence Thesauri, non-commercial use is the most common restriction (7 out of 8 thesauri). Moreover, ND restriction is often combined with NC restrictions (4 out of 5 thesauri forbid both).

5.3 Overall discussion and recommendations

The overall results of the reusability analysis is summarized in Table 5, whose columns refer to the three categories concerning 5 star LD evaluation while the rows refer to those identified for licence evaluation. We can observe that most of the thesauri with higher values (>= 4) for both 5 star LD principles and licence, (e.g., GEMET, AGROVOC, NERC NVS2.0, GBA, NALT and UNESCO) have been detected in more than one source of provenance in Table 1; this could imply that there is a direct relation between the "popularity" of a thesaurus and its "reusability". Moreover, the analysis performed on the thesauri in the catalogue shows an average good level of reusability. In fact, about the 58% of thesauri considered are Linked Data ready or RDF ready and are licenced with open or partially open licences. However, some recommendations to improve their reusability can be outlined:

- More attentions should be paid to HTTP dereferenceability of concepts URIs. Currently, Linked Data best practices seem quite popular among thesaurus providers in the environmental domain: about the 46% of the thesauri considered are already in Linked Data. However, the 54% of thesauri fails in a complete adoption of HTTP dereferenceable URI showing that HTTP dereferenceability is not yet received in the environmental thesauri community of providers. This shortcoming prevents the discovery and the integration of concepts from distinct thesauri in a follow-your-nose fashion hampering the jointly use of existing thesauri which is a requirement when managing geographical information at the European scale.
- Licence should be more carefully stated. More than 50% of the thesauri in the catalogue are released with licences from standard framework such as Creative Commons or equivalent. However, determining under which licence a dataset is released is still a time consuming activity. Depending on the thesaurus, the licence can be stated in different sources, e.g., the web site of the thesaurus, the web site of the institution owning the thesaurus, the related datahub page or related publications. Many thesauri are available in more than one of the aforementioned sources, but, rarely the licence is stated in all the sources available. In some cases, an explicit web link at the licence page is missing or it is not possible to find which version of the licence is adopted. As far as we have tested, generally no licence is included in the RDF returned by HTTP dereferencing.

6 Conclusion and future work

This paper provides a "reference catalogue" of thesauri available in the web for the environmental domain, in the perspective of the integration and the sharing of a large amount of existing environmental data provided by the National/Regional Environmental Agencies and other public and private environmental stakeholders. This is an emergent issue since several recent European directives address a more global management of environmental information in order to overcome cross-border/language and cultural barriers and to improve the cooperation between nations at European level. To this purpose, we present a methodology to identify terminological resources available on the web, possibly in Linked Data, a definition of domain independent criteria for the reusability based on two characteristics: the licence openness and the compliance to HTTP dereferenceability of URIs. Critical issues arising during the evaluation process are also detailed in the analysis. The future works will be twofold. On one side, we will complement the analysis presented considering notions of quality that have been recently proposed. In particular, multilingual support and SKOS-compliance of Linked Data and RDF ready thesauri can be analysed by using quality measures proposed in [15]. On the other side we will improve the dissemination of our results among the environmental communities developing a web portal to expose the whole catalogue and the reusability evaluation performed on each thesaurus.

Acknowledgements. The paper activity has been carried out within the EU funded project eENVplus (CIP-ICT-PSP grant No. 325232). The authors would like to thank all partners and, in particular, Paolo Plini (IIA-CNR) and Carlo Cipolloni (ISPRA) for their important collaboration. The authors would also like to thank the team of the European Commission's Joint Research Centre (Italy) for the valuable contribution.

References

1. Albertoni, R., Martino, M.D., Franco, S.D., Santis, V.D., Plini, P.: EARTh: An Environmental Application Reference Thesaurus in the Linked Open Data cloud. SWJ 5(2), 165–171 (2014)
2. Berners-Lee, T.: Linked data. `http://www.w3.org/DesignIssues/LinkedData.html` (2009), accessed: 20 March 2014
3. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J.: The AGROVOC linked datasetaset. SWJ 4(2), 341–348 (2012)
4. Carusone, A., Olivetta, L.: Italian Thesaurus of Earth Sciences (ThIST). APAT (2006)
5. De Martino, M., Albertoni, R.: A multilingual / multicultural semantic-based approach to improve data sharing in an SDI for nature conservation. Int. J. of Spatial Data Infrastructures Research 6, 206–233 (2011)
6. Faber, P.: A Cognitive Linguistics View of Terminology and Specialized Language. Walter de Gruyter (2012)
7. Government Linked Data Working Group: W3C Working Group Note: Best Practices for Publishing Linked Data. `http://www.w3.org/TR/ld-bp/` (2014), accessed: 24 March 2014
8. Grignetti, A., Plini, P., Mazzocchi, F., De Santis, V.: A thesaurus for remote sensing and gis: preliminary version and future plans. In: 19th Int. Conf. Informatics for Environmental Protection. pp. 783–787 (2005)
9. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool (2011)
10. Manaf, N., A., Bechhofer, S., Stevens, R., Manaf, N.: The current state of SKOS vocabularies on the web. In: 9th Int. Conf. on The Semantic Web: Research and Applications. pp. 270–284 (2012)
11. Miles, A., Bechhofer, S.: W3C Recommendation: Simple Knowledge Organization System Reference. `http://www.w3.org/TR/skos-reference` (2009), accessed: 20 March 2014
12. Palavitsinis, N., Manouselis, N.: A Survey of Knowledge Organization Systems in Environmental Sciences. In: Athanasiadis, I.N., Rizzoli, A.E., Mitkas, P.A., Gómez, J.M. (eds.) Information Technologies in Environmental Engineering, pp. 505–517 (2009)
13. Plini, P., Salvatori, R., Valt, M., De Santis, V.: SnowTerm: a terminology database on snow and ice. In: 21st Polar Libraries Colloquy. pp. 82–89 (2006)
14. Rodríguez-Doncel, V., Gómez-Pérez, A., Mihindukulasooriya, N.: Rights declaration in linked data. In: 4th Int. Work. on Consuming Linked Data (2013)
15. Suominen, O., Mader, C.: Assessing and improving the quality of SKOS vocabularies. J. on Data Semantics 3(1), 47–73 (2014)
16. Zapilko, B., Schaible, J., Mayr, P., Mathiak, B.: TheSoz: A SKOS representation of the thesaurus for the social sciences. SWJ 4(3), 257–263 (2013)