Accepted version for Earth Sci Inform (2018). https://doi.org/10.1007/s12145-018-0344-8

LusTRE: A Framework of Linked Environmental Thesauri for Metadata Management

Riccardo Albertoni, Monica De Martino, Paola Podestà
Institute for Applied Mathematics and Information Technologies,
National Research Council (CNR-IMATI),
Via de Marini 6, 16149 Genoa, Italy
firstname.lastname@ge.imati.cnr.it
monica.demartino@ge.imati.cnr.it, +39 0106475662
https://orcid.org/0000-0002-1963-3321

Andreas Abecker, Roman Wössner, Karsten Schnitter
Disy Informationssysteme GmbH
Karlsruhe, Germany
firstname.lastname@disy.net

Abstract The paper illustrates a Linked Thesaurus Framework for the Environment, named LusTRE, to facilitate data sharing across different environmental disciplines. It provides a knowledge infrastructure of multilingual thesauri and code lists, interlinking them so that they can be used as one integrated linked data source. This multilingual thesaurus is published according to the Linked Data Best Practices and supports metadata compilation and data discovery for describing and finding Environmental geodata. A human readable web interface is provided for the exploitation of LusTRE as well as a set of web services for programmatic access to the knowledge infrastructure.

LusTRE has been exploited within the European directive INSPIRE and SEIS piloting testbeds implemented within the EU project eENVplus in order to support cross-border and cross-domain data sharing. It is aimed at supporting multilingual data search and query refinement. In order to show how interlinked content can help users to more easily express metadata within Spatial Data Infrastructures (SDI), LusTRE web services have been integrated within existing metadata editors and geoportals.

Keywords SKOS, thesaurus, Linked Data, web service, metadata management, INSPIRE

Introduction

INSPIRE is the European Directive that commits EU member states to establish a Spatial Data Infrastructure (SDI) for Europe to enable the sharing of environmental spatial information among public-sector organizations and to better facilitate public access to spatial information across Europe (INSPIRE 2015a). INSPIRE addresses 34 data themes (such as cadastral parcels, hydrography, protected sites, agricultural and aquaculture facilities, sea regions, and natural risk zones) needed for environmental applications, with key components specified through technical implementing rules (INSPIRE 2015b). SEIS, the Shared Environmental Information System for Europe (http://ec.europa.eu/environment/archives/seis/), is aimed at improving access to environmental information for local and national governments, non-governmental organizations, academia, and the public.

The European CIP-PSP co-founded project eENVplus (Attardo and Saio 2013) developed a complete ecosystem of software services for supporting all aspects of interoperable environmental data on the Web. The software services were developed and tested in ten INSPIRE and SEIS testbeds for pilot implementation. The eENVplus testbeds (http://showcase.eenvplus.eu/client/) comprised, for instance, a cross-border pilot in Hungary and Slovakia regarding protected areas, a cross-border pilot in Italy and Slovenia on geological-map harmonization, a pilot in Belgium and another one in Italy on the implementation of SEIS for air quality data, together with a Greek pilot regarding decision support in forest-fire management.

According to the INSPIRE implementing rules (INSPIRE 2015b), metadata must be added for all the information published in an SDI. The core metadata profile of ISO 19115 (ISO 19115-1:2014) is adopted to describe general aspects of the actual data such as, the creator, geographical coverage, or subject matter. In order to "standardize" the content descriptions of published data as a defined framework of the INSPIRE Directive, keywords are normally taken from pre-defined thesauri, taxonomies, code lists or domain ontologies. Such instruments for organizing information and promoting knowledge management are called Knowledge Organization Systems (KOSs).

In this paper, we focus mainly on thesauri, which are the most common KOSs for metadata annotation and data search. They are a type of highly structured vocabularies characterised by a list of concepts that are hierarchically organised into a network structure. We also consider various code lists, i.e. a kind of flat vocabulary characterised by a list of verbal descriptions of things and events with numerical or alphabetical codes used in a specific area (Manaf 2012). In the paper, we use the term "vocabulary" to refer to both a thesaurus and/or code list.

Thesauri are often used by domain experts and typically cover only narrow knowledge areas (but very extensively). They are generally not translated into all European languages and, naturally, in order to structure their concepts, they address the community-specific perspective they were built for. Different thesauri are thus employed to describe overlapping, related or even identical topic areas, or covering different environmental disciplines. However, the diverse views of reality impede the sharing of information across different communities, which may limit the interoperability of the metadata expressed in these different thesauri. To manage the different thesauri is needed.

The aim is to enable users to walk, i.e. cross, easily from one thesaurus to the next. The challenge is to exploit all these pieces of valuable information in a consistent and integrated way in order to support cross-organizational, cross-disciplinary and cross-language metadata management (for instance, when searching for all data regarding one specific entity in the world, such as a specific biotope, data coming from different data sources that have been annotated with respect to different thesauri). To create a common thesaurus covering the diverse views of the environmental domain is neither realistic, nor desirable because all the different views are useful. Thus, there should be discipline-specific thesauri and the responsibility for their maintenance should remain with their creators.

In this paper, we provide a solution for the integration and joint exploitation of environmental thesauri promoting a Linked Thesaurus Framework for the Environment, named LusTRE, which we developed within the two consecutive EU funding research, development and pilot projects, NatureSDIplus (http://www.nature-sdi.eu/) and eENVplus (http://www.eenvplus.eu/). The framework provides a system for connecting environmental thesauri also including a few code lists (as an additional requirement of the project community) into one multi-vocabulary. The Simple Knowledge Organization System (SKOS) developed by the World Wide Web Consortium (Miles and Bechhofer 2014), and the Linked Data Best Practices promoted by Berners-Lee in (Berners-Lee 2009) are employed for encoding, sharing and linking the vocabularies on the Web.

The main contributions of LusTRE are: (i) a unique point of access to several environmental thesauri and code lists which can be enriched by linking new thesauri; (ii) a high degree of reusability in terms of easy access and exploiting its content (human-readable web pages, web services, SPARQL end points); (iii) a full and transparent exploitation of the interlinking of thesauri. Thus, a query involving a concept returns all the information related to it (e.g. preferred label, alternative label, relationships) and provided by equivalent concepts in the interlinked vocabularies.

In the paper, we present the design and architecture of LusTRE, its major components, content and web services, and how to exploit it within client applications for metadata management. The cross-walking feature is one of the web services provided in LusTRE for the full transparent exploitation of the interlinking among the vocabularies for the user. Cross-walking means: if identical, similar or related concepts in different thesauri are connected by one (or more) link(s) in LusTRE, an application may start from one specific label / concept in one thesaurus, walk from this thesaurus to another one (or to more than one) through the link(s) and thus consider for the given application purposes, not only the knowledge contained in the starting thesaurus, but also that contained in the linked thesauri. Thus, different (natural or technical) languages, multiple viewpoints from different disciplines, or different modelling methodologies (deeper or shallower, broader or narrower, different structuring principles) can be exploited at the same time.

We illustrate two main contributions provided by the connections among thesauri or code lists (aka. linksets) and the web services used to exploit them: (i) a multilingual enrichment of a thesaurus in terms of new translated labels by navigating its connected thesaurus through the linkset by exploiting a cross-walking web service. This eases the incomplete language coverage issue affecting many popular SKOS thesauri, which are provided in all the expected languages only for a subset of their concepts; (ii) query refinement. Through the cross-walking services, users may extend the number of new concepts reached by navigating a set of links, thus enlarging the space of concepts that can be browsed (aka, the thesaurus browsing space) and used to refine the query. LusTRE also offers a visual exploration tool, which provides a compact view of the thesaurus browsing space.

The paper is structured as follows. “A review of related thesauri and tools” provides an overview of the thesauri available and the tools used to manage them, “Design issues” provides the design of the framework and an overview of the system components and their interplay. The components are described extensively in the sections “The LusTRE-VOC knowledge infrastructure”, “The LusTRE-ES exploitation services” and “Web interface and the LusTRE-WEBe exploration tool”. Before the “Discussion and Conclusions” section, we show how LusTRE can be applied for metadata management in “LusTRE application for metadata management” and how to access to LusTRE in “LusTRE access”.

A review of related thesauri and tools

The development of an SDI at a European level requires the deployment of geographical data in a standardized way and with common nomenclatures.

Why thesauri? Thesauri enable communication among the various communities working on environment-related disciplines (e.g., chemistry, geology, and biology). They allow users to share and agree on scientific/technical terms, and to express them in multiple natural languages. Different communities with a large spectrum of competencies are involved in the treatment and management of geographical information. Thus, SDIs need to deal with several thesauri and with code lists as common nomenclatures in order to cover such a large spectrum of competencies.

Their application is crucial in semantic interoperability in order to improve the result of data access and integration (e.g., metadata editors and geoportals, semantic annotation, automated tagging datasets). For example, Wright et al. (2015) design the CAST thesaurus for tagging chemistry datasets. Da Silva et al. (2009) propose using thesauri to generate Indexed Base with terms, relationship and occurrence between terms and resources in order to minimize metadata recovery problems in SDIs. Maué and Ortmann (2009) highlight the need to be able to move from one information community to another. They propose semantic annotations to facilitate the discovery and evaluation of geographical information. They do not put forward a specific approach but argue how semantic web technologies can help to create an SDI for the Amazon, which facilitates the involvement of economic, ecological and ethological information communities of.

Simple Knowledge Organization Systems (SKOS). In 2009, the World Wide Web Consortium (W3C) established the SKOS standard model for expressing the basic structure and content of thesauri and other similar types of systems for knowledge organization (Miles and Bechhofer 2009). SKOS thesauri are expressed as RDF (Resource Description Framework) triples and can be published according to Linked Data principles. SKOS provides lightweight semantics for terminology concepts (Miles and Bechhofer 2009). Linked data allow concepts to be composed and published on the World Wide Web, linked with data sources on the Web and integrated into other concept schemes (Hyland et al. 2014; Heath and Bizer 2011). Relevant properties can be expressed, such as concepts identified using URIs, labelled with lexical strings in one or more natural languages (see *skos:prefLabel*, *skos:altLabel*), linked to other concepts and organized into informal hierarchies and association networks (e.g., *skos:broader* and *skos:narrower* properties define hierarchical links, *skos:related* associates two SKOS concepts). The semantic relations above are usually aggregated into concept schemes. SKOS enable mappings to be expressed between concepts that belong to separate concept schemes. Concepts belonging to separate schemes might be equivalent (*skos:exactMatch, skos:closeMatch*), more specific (*skos:broadMatch),* less specific *(skos:narrowMatch*), or related (*skos:relatedMatch*).

According to Baker et al. (2013), *"skos:closeMatch* was intended for use with concepts that are sufficiently similar to be used interchangeably in a given context. The property is not defined as transitive in order to avoid the uncontrolled propagation of the similarity relation to further contexts. *skos:exactMatch*, defined as a transitive sub-property of *skos:closeMatch*, was intended to express a degree of similarity close enough to justify such a propagation. The mapping properties *skos:broadMatch*, *skos:narrowMatch*, and *skos:relatedMatch* are the sub-properties, respectively, of *skos:broader*, *skos:narrower,* and *skos:related*".

Thesauri and code lists for the environment. Several thesauri and code lists for the environment have been provided as RDF/SKOS for consultation and access to content on the Web. The general Multilingual Environmental Thesaurus GEMET (EIONET 2015) is the official thesaurus for the SDI in the European Community promoted by the INSPIRE directive. The INSPIRE infrastructure also includes a registry (INSPIRE 2015c), a public reference directory including a set of INSPIRE code lists. AGROVOC (http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus) is the multilingual agricultural thesaurus covering all areas of interest of the Food and Agriculture Organization of the United Nations, EARTh is the bilingual Environmental Applications Reference Thesaurus (Mazzocchi and Plini 2005) promoted by the Italian National Research Council CNR-IIA. ThiST is the bilingual Italian Thesaurus of Sciences of the Earth (Carusone and Olivetta 2006) made available by the Geological Survey of Italy in ISPRA. SoilThes (https://goo.gl/oA2L1m) is the General Multilingual Soil Thesaurus created in the eContentplus project GS SOIL (GS SOIL 2015) as an extension of GEMET with soil vocabulary of ISO 11074 and addition soil-specific concepts (L'Abate et al. 2015).

Many of these thesauri are published as Linked Data (Albertoni et al. 2014b) and are interlinked with each other. For example, GEMET is mapped to AGROVOC (Caracciolo et al. 2015) and EARTh (Albertoni et al. 2014a) is mapped to AGROVOC, GEMET, EuroVoc (http://eurovoc.europa.eu/drupal/).

Framework/Tool. The provision of a single point of access to these SKOS vocabularies is challenging and is addressed by two types of frameworks: knowledge bases addressing specific disciplines and tools to manage the vocabularies.

Examples of knowledge bases for target disciplines are those proposed by Schreiber et al. (2008) for the cultural heritage; Wolstencroft et al. (2011) for biology; (http://eurovoc.europa.eu/drupal/) for the activities of European Union; EIONET (2015) and Albertoni et al. (2014.a) for the environment; Caracciolo et al. (2013) and Baker et al. (2016) for agriculture; and L'Abate et al. (2015) for soil science.

Examples of web application tools for managing SKOS vocabularies include: VocBench (Stellato et al. 2015) a collaborative web-based multilingual tool for thesaurus editor, validation and publication and its new version VB3 (http://aims.fao.org/activity/blog/vocbench-3-free-and-open-source-platform-editing-ontologies-thesauri-and-rdf-datasets) improving the editing capabilities with facilities for management of OWL ontologies and SKOS/SKOS-XL thesauri (Stellato et al. 2017); PoolParty (https://www.poolparty.biz/) (Schandl et al. 2010) a proprietary semantic middleware platform providing a web-based thesaurus editor using Linked Data; and TemaTres (http://www.vocabularyserver.com/index.html) an open source web application for the management, publication and sharing of controlled vocabularies. These tools provide technologies for the publication, semantic interlinking and access of vocabularies, but none of them provide content.

LusTRE does not aim to be a management tool like PoolParty or VocBench but focuses more on the knowledge base, facilitating the joint exploitation of the most common vocabularies for the Environment. With respect to AGROVOC (Caracciolo et al. 2013), GEMET (ELIONET 2015) and EuroVoc, LusTRE has several advantages. AGROVOC, GEMET and EuroVoc provide an overall thesaurus of different linked thesauri, and web services to access the information associated with a concept, but they do not exploit interlinked thesauri, and they also provide web services that given a URI return the corresponding URIs (*skos:exactMatch, skos:closeMatch, skos:broadMatch or skos:narrowMatch*).

LusTRE manages and exploits each thesaurus individually and promotes the joint exploitation of thesauri, building a more exhaustive knowledge base for the Environment. It makes a few vocabularies that have not yet been published available as Linked Data (e.g., EARTh and ThiST in the LOD Cloud) and provides a link among them or those already available in Linked Data. It thus becomes a unique point of access to several vocabularies for the Environment. It also provides a high degree of reusability, in terms of being easy to access and exploit its content (human readable web pages, web services, SPARQL end points). It exploits interlinking among thesauri in a transparent way, thus, a query involving a concept returns all the information related to it and at the same time, the information on related concepts in other thesauri.

LusTRE design

This section outlines the design principles of the LusTRE framework and its software components.

Issues and objectives

The reusability of thesauri and code lists for the Environment is fundamental within an SDI in order to provide homogeneity in data description and data discovery. The issue is how to exploit their heterogeneity.

Our objective is to create an open environment where all these vocabularies can be available to support better metadata compilation and data discovery for describing and finding data and services. The main goal of such a framework is to preserve and retrieve the information based on term definitions, synonyms, and semantic relations, rather than just keywords. This would guarantee the uniformity of the persisted metadata information, as well as the discoverability of metadata based on the semantic meanings even if metadata include diverse and dissimilar keywords.

We have built a framework that enables existing environmental vocabularies to be combined. It considers the heterogeneity in scope and levels of abstraction of environmental terminologies as an asset when managing environmental data, thus exploiting Linked Data best practices (Hyland et al. 2014; Heath & Bizer 2011) to provide a multi-thesaurus solution for INSPIRE data themes related to the environment.

Design requirements and software components

LusTRE was developed by extending the common thesaurus framework for Nature Conservation described in De Martino and Albertoni (2011) resulting from the previous NatureSDIplus project. LusTRE provides the interlinking of different thesauri and code lists in order to achieve a broader domain coverage, more multilingualism, and novel cross-discipline exploitation services. The framework is a stand-alone software infrastructure, which can be exploited in a service-oriented manner by other software tools such as metadata editors or catalogue services.

It is designed to be an open and dynamic environment where it is possible to add, extend, assemble and share general-purpose thesauri as well as domain-specific vocabularies. It thus has to have the following design requirements:

- *Modularity*. Each vocabulary is a module that is plugged into the framework. Modularity is preserved in order to include future updates for existing vocabularies.
- *Openness.* Each vocabulary is easily extendable, in order to add (as separated modules) new concepts and terms keeping the original vocabularies separate.
- *Exploitability*. Each vocabulary is encoded in a standard and flexible format, in order to encourage the adoption and the enrichment by third party systems.
- *Interlinking*. Terms and concepts in existing vocabularies are interlinked in order to harmonize the use of terms from a multicultural/multilingual point of view.

Semantic Web technology is used to meet these requirements. Hence, thesauri and code lists are encoded in SKOS and represented in the RDF data model. Linked Data principles are employed to expose, share, and connect (interlink) vocabularies via Uniform Resource Identifiers (URI) that can be looked up on the Web. The encoding to SKOS enables modularity, whereas accessibility via linked data ensures openness and exploitability.

We thus decided to build a Linked Data solution based upon state-of-the-art Semantic Web technologies, namely the LusTRE-VOC knowledge infrastructure depicted in Fig. 1.

LusTRE-VOC is a knowledge base, which includes a set of thesauri and code lists published as linked data as well as any links between them.

Technically, LusTRE-VOC is based on Virtuoso (https://virtuoso.openlinksw.com/), which is a multi-purpose and multi-protocol data server, which can be used to store and query RDF knowledge bases. RDF is the representation language of the Semantic Web and easily represents arbitrary knowledge structures. Virtuoso offers a SPARQL endpoint. SPARQL (http://www.w3.org/TR/rdf-sparql-query/) is the W3C standard query language for RDF.

Based on this knowledge infrastructure, the following further requirements were addressed in order to allow for optimum usability and maximum usefulness of the knowledge contained in LusTRE:

- Accessibility by third party tools. The knowledge contained in LusTRE-VOC can be exploited by many existing tools, without the need to change technologies and move to a solution for metadata management.
- Accessibility by a Direct User Interface. For a human user with a web browser, an *easy*-to-use web interface is provided to explore, understand and use the knowledge represented in LusTRE-VOC.

The first requirement led to the design and implementation of the LusTRE Exploitation Services (LusTRE-ES in Fig. 1). These are web services, which can be called through a REST interface, embeddable into any existing tool that does not require the end user to know anything about the underlying semantic technologies, but instead, offers a number of useful services for intelligently dealing with metadata keywords and concepts.

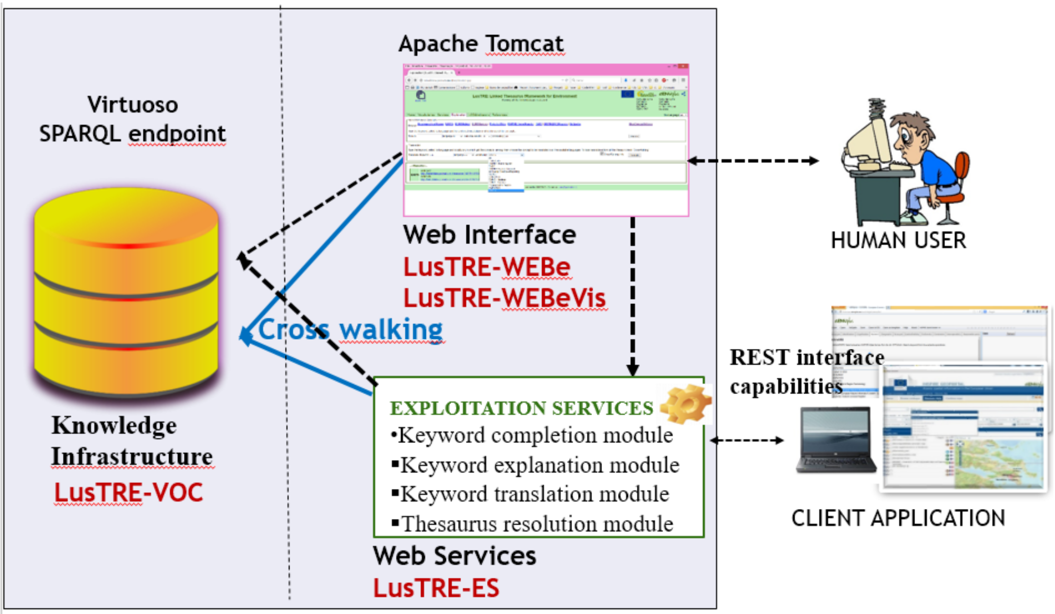


Fig. 1 LusTRE component connections

The service manager delegates the execution of each service invocation to a service module (sometimes several service modules may be suitable for a given user request). The service module translates the service call into SPARQL queries and hands the queries over to the Virtuoso server, collects the outcomes, if necessary merges them, and delivers the result to the client. During the execution of the SPARQL query by the Virtuoso server, different vocabularies may be navigated, thus connecting different knowledge domains or bridging between different languages, etc.

The second requirement led to the implementation of the LusTRE web interface in Fig. 1. This makes available: (i) the explorative tool LusTRE-WEBe, which provides a human-accessible interface to manually search and browse the knowledge infrastructure or by exploiting several LusTRE-ES through a text-based browser; (ii) the visual browser LusTRE-WEBeVis which provides a graphical interaction with the terminological knowledge and illustrates the knowledge graphs and their interdependencies.

The three LusTRE software components and their connections are illustrated in Fig. 1.

LusTRE-ES services are used in client applications such as metadata editors, catalogue services, or geoportals to access the knowledge base in LusTRE-VOC. The joint use of the vocabularies within LusTRE relies on newly provided connections (aka, linksets) among terminologies and the notion of "cross-walking" for the exploitation of such connections. Cross-walking enables the user to automatically navigate among matching concepts belonging to the different linked vocabularies, thus to get information that is missing in one but available in others, and thus to discover additional information for a specific concept. It facilitates the joint exploitation of different vocabularies working beyond the scope and limitations of a single one, possibly enriching the data at hand, and, thus, improving user satisfaction.

In addition, LusTRE provides the stand-alone use of a vocabulary, which is the usual practice, similar to what is provided by the websites of some thesauri such as GEMET and AGROVOC (Caracciolo et al. 2013). LusTRE's contribution in terms of the stand-alone use of a vocabulary is the publication as linked data. Thus thesauri that are not yet published in the LOD, such as ThIST and EARTh (Albertoni et al. 2014a), have been included in the Linked Open Data cloud (http://lod-cloud.net/).

The three components are described individually in the following sections.

The LusTRE-VOC knowledge infrastructure

The LusTRE-VOC knowledge infrastructure is accessible as one integrated linked data source. It provides the following features:

1. Stores and provides domain thesauri and code lists (i.e. stores the SKOS data records for the concepts of a thesaurus).
2. References the domain thesauri and code list concepts using dereferenceable URIs.
3. Represents the interlink between thesaurus/code lists (i.e. stores thesaurus linksets, which are the set of links that establishes typed connections between concepts of one thesaurus to concepts of another).

The third features is the most interesting because it enables one conceptual network to be woven from a number of formerly separated vocabularies, thus avoiding the limitations of a single vocabulary in terms of scope, breadth, depth, multilingualism, etc.

The SKOS representations of several existing vocabularies have been created and stored in LusTRE-VOC. If a vocabulary is already published as SKOS following the LOD principles, it remains with the creator and only the interlinking to external sources is provided in LusTRE-VOC.

LusTRE-VOC implementation methodology

The following multi-step process was applied to build LusTRE-VOC:

1. **Vocabulary selection**. Existing environmental thesauri and code lists for the Environment were identified. We classified them using reusability criteria (license openness and compliance with Linked Data best practices), and also considering coverage of the INSPIRE data themes. The analysis considers the following: (i) the use of standard technological and scientific terms in relation to geographical data; (ii) the exploitation by different users operating in the fields of the INSPIRE data themes; (iii) the level of compliance with some reusability criteria based on the 5-star Linked Data principles (Berners-Lee 2009) and license openness. As described in Albertoni et al. (2014b), we assessed the reusability criterion by considering the *licence openness* and the compliance with LD focusing on the *usage of dereferenceable HTTP URIs* as identifiers for concepts. Licence and HTTP dereferenceability are central prerequisites for every reuse scenario and they are crucial for interlinking structured data (Berners-Lee 2009). Concerning the INSPIRE data theme coverage, domain thesauri or code lists were included in LusTRE (e.g., EUNIS species and EUNIS habitat for the Habitats and Biotopes data themes). Besides these specific vocabularies, more general-purpose thesauri for the environmental and geographical field are also plugged into LusTRE as trans-theme thesauri (e.g., EARTh, GEMET, AGROVOC, EuroVoc).
2. **Vocabulary processing.** The selected vocabularies were encoded in SKOS format and inserted in LusTRE. Among the selected vocabularies to be inserted in LusTRE, we identified those not natively available as RDF, or with some errors (e.g., inconsistent encodings) and with structural problems. We then used LODrefine (http://code.zemanta.com/sparkica/) to clean up and encode the vocabularies in SKOS/RDF.
3. **Vocabulary interlinking**. Two-way links among vocabularies inside LusTRE and to external LD thesauri were generated. Among the vocabularies inserted in LusTRE, the connections were first automatically generated using various tools and then validated by experts from the specific domains of the respective thesauri. Firstly, SILK (http://silk-framework.com/) was applied to discover new links, then the SILK results were validated by some of the domain expert communities in order to verify the accuracy of the links and to identify the most suitable types of interlinking properties (e.g., to distinguish between *skos:exactMatch* and *skos:closeMatch*). Further equivalences were created by materializing the inverse and the transitive closure on *skos:exactMatch*. For example, with the equivalences from EARTh to GEMET, and from GEMET to AGROVOC, EuroVoc, DBpedia and UMTHES, we generated the EARTh outgoing links to AGROVOC, EuroVoc, DBpedia and UMTHES traversing GEMET's links (Albertoni et al. 2014a).
4. **Vocabularies and interlinking publication**. Vocabularies not yet published as LD and the linksets are made available in LusTRE-VOC in accordance with Linked Data Best Practices. Virtuoso and Apache Tomcat were set up to improve LusTRE scalability on the web. Virtuoso was adopted as the RDF Store to store the various vocabularies in the knowledge infrastructure. Virtuoso also provides a SPARQL endpoint that can be used by Semantic Web literate users in order to perform SPARQL queries on LusTRE content.

The same methodology can be deployed to add a new vocabulary in LusTRE. In particular, if the vocabulary is not already presented according to Linked Data best practices, its inclusion in LusTRE consists of steps 2-4 above. Otherwise, if the vocabulary is already available in the LOD, only the interlinks need to be generated and published.

Table 1 Thesauri and code lists presented in LusTRE

Thesauri and Code Lists with new HTTP dereferenceable concept URIs and exposed in LusTRE's SPARQL endpoint			
Name	#Concepts	Languages (N of languages)	Owner
Thesauri			
EARTh	14,350	it en (2)	CNR-IIA
ThiST	34,150	it en (2)	ISPRA (IT)
EUNIS-Species	202,531	la (1)	ETC/BD for EEA and Eionet
EUNIS-Habitat	5,431	en es sv de el fi pt da fr it nl (11)	ETC/BD for EEA and Eionet
Code List			
EEA Biogeographical Region (DMEER)	81	en (1)	EEA
IUCN Protected Sites	9	en (1)	IUCN
INSPIRE_Registry-DataTheme	38	en bg sv es sl sk ro pt pl no mt lt lv it hu el de fr fi et nl da cs hr ca (25)	EC-JRC
INSPIRE_Registry-FeatureConcept	367	en (1)	EC-JRC
Eionet: Air Quality	886	en (1)	ETC/ACM and EEA

Table 2 Thesauri and code lists already available in LD and linked with LusTRE

Thesauri and code lists maintaining their original concept URIs			
Name	#Concepts	Languages (list/#)	Owner
Thesauri presented in LusTRE SPARQL endpoint			
GEMET	5,223	ar bg ca cs de el en-us en es et eu fi fr ga hr hu it lt lv mt nl no pl pt ro ru sk sl sv tr uk zh-cn (32)	EEA
AGROVOC	32,310	hu tr ar cs de en es fa fr hi it ja lo pl pt ru sk th zh te ko ms uk sv (24)	AIMS-FAO
EuroVoc	6,883	bg es cs da de et el en fr hr it lv lt hu mt nl pl pt ro sk sl fi sv mk sq (23)	EU commission
Additional vocabularies linked by LusTRE			
DBpedia	3 billion information (RDF triples)	ar cs de el es fr it ja ko nl pl pt (12)	crowd-sourced community
UMTHES	300,000	de en (2)	Federal Environment Agency, Germany
IUGS-CGI Code lists	Not available	21	Geoscience Terminology Working Group (GTWG)

Table 3 Number of links between thesauri in the rows and the thesauri in the columns. For each pair of thesauri, the first and second rows indicate the number of *skos:exactMatch* and *skos:closeMatch,* respectively

Mappings between Thesauri: Number of links for skos linksets (filter >200)							
	EARTh	ThIST	AGROVOC	GEMET	EuroVoc	DBpedia	UMTHES
EARTh		1140	1425	4328	1346		
		1140	1425	4328	1346	1862	2970
ThIST	1140		1695	792	792	733	921
	1140		1741	835	835	797	948
AGROVOC	1425	1695		1175	1269		
	1445	1741		1181	1269	11014	
GEMET	4328	792	1175		1683		
	4328	835	1175		1683	2035	3482
EuroVoc	1346	733	1269	1683			
	1346	797	1269	1683			

Content description

LusTRE-VOC includes 302,259 concepts, 1,467,416 terms, and 112,471 links. Tables 1 and 2 detail the characteristics of the vocabularies in LusTRE: number of concepts, languages, and original providers.

Table 1 shows the new linked data vocabularies for which we provide HTTP dereferenceable URIs. Table 2 shows the well-known vocabularies we included in LusTRE SPARQL endpoint, maintaining their original URI and other interlinked vocabularies.

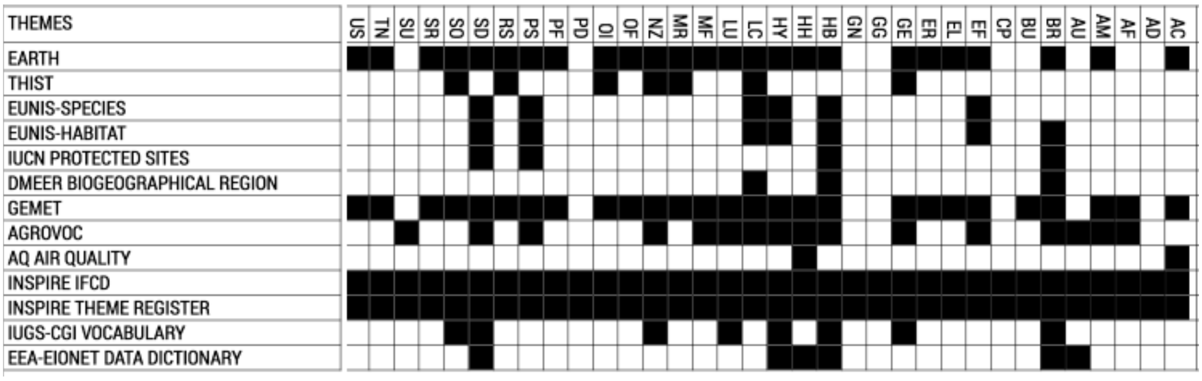
Table 3 lists the linksets between LusTRE vocabularies with more than 200 links and specifies the number of SKOS mappings expressed with the properties *skos:exactMatch* and *skos:closeMatch*.

In addition to the linksets in Table 3, LusTRE provides other mappings: 24464 *skos:relatedMatch* between EUNIS-Species and EUNIS-Habitats, 917 between GEMET and DBpedia, 217 *skos:broadMatch* between GEMET and EuroVoc, and 365 *skos:broadMatch* between INSPIRE Feature Concepts and INSPIRE Theme Register. Caching well-known vocabularies in the LusTRE SPARQL endpoint takes full advantage of the links to them, thus avoiding issues with federated queries (e.g., unresponsive or unavailable third party servers).

Figure 2 shows the distribution of LusTRE vocabularies (preferred labels) in relation to the INSPIRE data themes. Each row represents a vocabulary, and each column refers to a data theme (shortened name according to the INSPIRE Data Specification). A black square in a specific row and column means that the vocabulary of that row provides concepts for the data theme in the corresponding column.

It should be noted that LusTRE may be used to describe almost all INSPIRE data themes: the two general-purpose thesauri GEMET and EARTh are included and cover almost all themes. Domain specific vocabularies are available for the following data themes: PS (Protected Sites), BR (Biogeographical Regions), HB (Habitats & Biotopes), SD (Species Distribution), EF (Environmental monitoring facilities), SO (Soil), AQ (Air Quality).

Figure 3 shows LusTRE language coverage. LusTRE knowledge infrastructure covers about forty languages. The following histograms illustrate the overall distribution of LusTRE concepts (preferred label) considering the various languages. The red line represents the number of concepts solely provided by GEMET: it provides 5233 concepts, which are in almost all the languages. The sum of the number of concepts provided by combining GEMET with other vocabularies outperforms the number of those provided by GEMET (e.g., more than 60,000 concepts in English and in Italian, more than 30,000 concepts for Czech, German, French, Portuguese, Japanese and Chinese).



THEMES	US	TN	SU	SR	SO	SD	RS	PS	PF	PD	OI	OF	NZ	MR	MF	LU	LC	HY	HH	HB	GN	GG	GE	ER	EL	EF	CP	BU	BR	AU	AM	AF	AD	AC
EARTH	■	■		■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■			■	■	■	■			■		■			■
THIST					■		■				■		■	■			■						■											
EUNIS-SPECIES						■		■									■	■		■						■								
EUNIS-HABITAT						■		■									■	■		■						■			■					
IUCN PROTECTED SITES						■		■												■									■					
DMEER BIOGEOGRAPHICAL REGION																	■			■									■					
GEMET	■	■		■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■			■	■	■	■		■	■		■	■		■
AGROVOC			■			■		■					■		■	■	■	■	■	■			■			■			■	■	■	■		
AQ AIR QUALITY																			■															■
INSPIRE IFCD	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
INSPIRE THEME REGISTER	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
IUGS-CGI VOCABULARY					■	■							■			■		■		■			■						■					
EEA-EIONET DATA DICTIONARY						■												■	■	■									■	■				

Fig. 2 Preferred labels of vocabularies (rows) and their coverage with respect to the different data themes (columns)

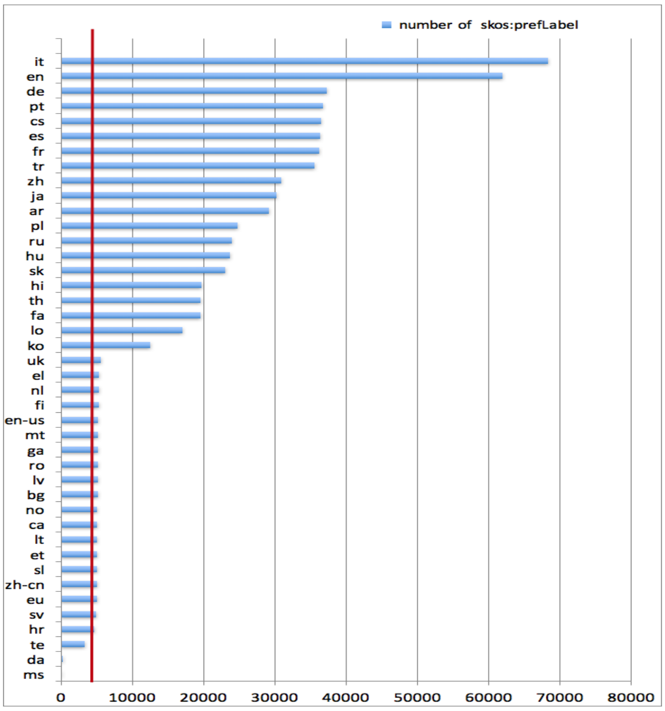


Fig. 3 Preferred-Label distribution with respect to the languages

The LusTRE-ES exploitation services

A set of web services was implemented to exploit LusTRE-VOC in third party applications. This section outline the software architecture of LusTRE web services. The implemented services are listed. One specific service is then described in more detail.

Architecture of LusTRE exploitation server

Figure 4 illustrates the overall idea of how to make use of the LusTRE knowledge base for improving metadata compilation and data discovery. On the one hand, there are end users who create, edit or search for metadata in order to describe or use spatial data sources or spatial data services. They employ different kinds of metadata-based tools (such as metadata editors, CSW tools, geodata portals). LusTRE offers highly configurable web services with standard-compliant interfaces to enable users to extend their existing metadata tools with functionalities that exploit the LusTRE knowledge. This means that, a "client" in Fig. 4 could be any tool for creating, processing or searching for metadata. On the other, there are different vocabularies, which via LusTRE interlinking mechanisms are bundled into a large, multilingual and multidisciplinary conceptual space that is represented in the Virtuoso store and can be accessed directly through Virtuoso's SPARQL endpoint.

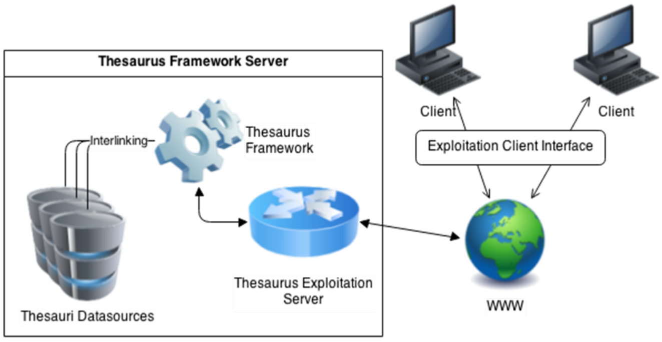

Thesaurus Framework Server
Interlinking
Thesaurus Framework
Thesauri Datasources
Thesaurus Exploitation Server
Client
Client
Exploitation Client Interface
WWW

Fig. 4 Usage scenario for interlinked vocabularies through exploitation services

Instead of programming against this SPARQL interface (which is, of course, possible), the LusTRE Exploitation Server offers a less technology-oriented and more end-usage oriented interface which manages all interactions with clients and abstracts away from the details of the SPARQL queries. In order to ensure expandability and flexibility, the LusTRE Exploitation Server is designed. It consists of: (i) an HTTP Server that offers a REST interface; (ii) a Service Manager; and (iii) a non-empty set of Service Modules.

The entry point for any call to the exploitation services is the *HTTP Server* with its REST endpoint. It decodes the HTTP request and creates a suitable query to the Service Manager.

The Service Manager delegates requests from the HTTP server to all suitable Service Modules (identified by their capabilities). The Service Modules return executable units of work (known as Callables) that will retrieve the query responses from the SPARQL endpoint. The Service Manager orchestrates the execution of the Callables and merges their results. It then routes back the merged query result to the REST endpoint / HTTP server. There, a JSON response is created and sent back with the HTTP response.

The Service Modules implement the logic of various functionalities in order to exploit the terminological knowledge. Specifically, they translate an end-user request into a (set of) SPARQL query(ies). They are the only points of contact between the thesaurus framework back-end and the rest of the exploitation layer.

The modules are pluggable cartridges for the exploitation server, each of which enhances the behaviour of the system in certain domains or with respect to certain criteria.

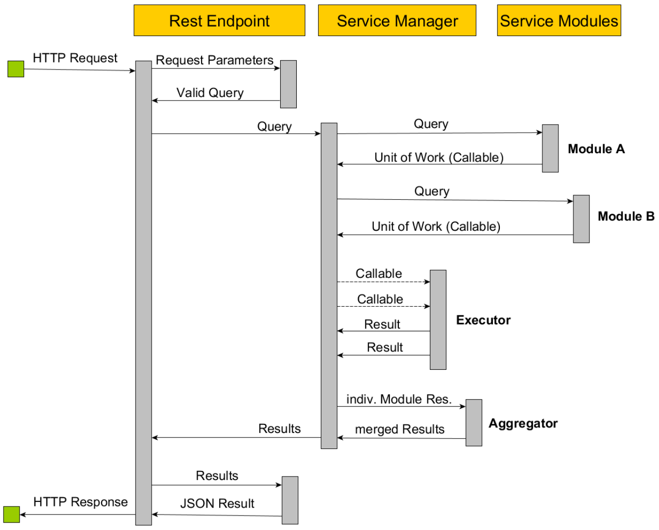


Fig. 5 Technical details of web services diagram

The connection between components and source projects is illustrated in Fig. 5, which depicts a sequence diagram. The architecture of the system enables the service modules to be independent of each other, and to focus on the one and only task they have: "Perform a specialized query from the thesaurus framework".

For each request type supported by a Service Module, a separate interface that has a public method must be offered, which returns a *Callable<T>* as a result. Each Service Module can be developed in a single class which must implement at least (but not limited to) one of the request-type interfaces. It must include an implementation of all the public fields and methods of the request-type interface. The rest of the logic, as well as any number of private methods and fields, can be freely defined and developed by the programmer in order to address the task that the service module has to carry out.

Following the LusTRE design guidelines for Service Modules, a uniform interface for all the Service Modules can be maintained. This means that every Service Module includes *only* the application logic for the task, and the Service Manager can orchestrate the execution of the processes. The entry point for any call to the service is the corresponding REST endpoint. This decodes the HTTP request and creates a suitable query to the service manager. The service manager delegates the query to all suitable modules. The modules return executable units of work that will retrieve the query responses from the SPARQL endpoint. The service manager executes these units of work. After merging the results of the different executions, the service manager returns the results to the REST endpoint. After creating the JSON response, the JSON response is sent with the HTTP response. Note that the Service Modules do not retrieve any results from them. They only provide units of work to be executed by the Service Manager. Hence, the control of the actual retrieval resides with the Service Manager.

LusTRE-ES was developed as a Java Web Servlet based on the Spring framework (http://projects.spring.io/spring-framework/). Spring is used as a dependency injection container. The discovery and configuration of Service Modules are also managed by this framework. The REST interface is implemented using Spring WebMVC. Any incoming request is thus handled by the DispatcherServlet of Spring, which routes it to a corresponding endpoint depending on the semantics of the request. Spring also marshals the results into a JSON response. The core functionality of LusTRE-ES is the mapping of the HTTP requests to appropriate SPARQL requests, which are then sent to the SPARQL endpoint. For this mapping, a custom Java DSL was developed around the Apache ARQ framework (https://jena.apache.org/). This DSL makes common tasks in the queries easier, e.g., the formulation of a language condition for queried labels or the determination of exact matches.

The ARQ framework is also used to make requests to the SPARQL endpoint, and manages its own pool of HTTP-client connections. This procedure is completely transparent within LusTRE-ES and no special configuration is used at present. The LusTRE-ES code is available through GITHUB

Table 4 Parameters of the GetSuggestions operation

Parameter	Type	Mandatory	Description
keyword	String	required	Keyword as search criterion, beginning of label to be searched for.
maxCount	Integer	optional	Maximum number of results.
thesauri	String	optional	Comma-separated list of vocabulary-URIs. Default: all.
service	String	optional	Comma-separated list of service as given by GetCapabilities.
languages	String	optional	Comma-separated list of languages based on ISO-639-1. Default: all.
cross-walking	Boolean	optional	Default is false, see below.
source	Boolean	optional	Default is false, when the true result contains a vocabulary source.

(https://github.com/eENVplus/tf-exploitation-server). The application can be deployed in any web application server, such as Apache Tomcat.

The set of exploitation services in LusTRE

The following services are offered to client applications through the REST interface:

- GetCapabilities describes the services offered by LusTRE-ES. The response contains (1) service metadata; (2) operation descriptions for the implemented service modules, their operations and the parameters of these operations; (3) descriptions of the terminologies exploited by this instance of LusTRE-ES.
- DescribeConcept, for a given concept URI, returns the SKOS descriptions from the associated terminologies by directly handing over a SPARQL request with the given URI to the Virtuoso instance that stores the associated terminological knowledge.
- GetSuggestions: if the user starts to type some keywords and the first few keystrokes are sent to the LusTRE-ES, this operation offers potential keyword completions.
- GetRelatives retrieves all concepts related to the input concept through a broader-concept, related-concept or narrower-concept relation. If specified by the cross-walking parameter, it also takes into account inter-vocabulary links.
- GetSynonyms finds all synonymous keywords for a given concept. If specified by the cross-walking parameter, it also takes into account the links between vocabularies.
- GetTopMostConcepts returns the top-level concepts of a given vocabulary.
- ResolveThesaurus finds the vocabulary it is contained in for a given concept.
- Visualization, for a given concept, retrieves the semantic information, interlinking information, as well as information on its surrounding concepts in the conceptual space described by the linked vocabularies. The retrieved information is provided as input for a graphical representation, such as a cluster visualization, implemented using the D3 JavaScript library.
- SPARQL: allows users to run their own SPARQL queries against the Virtuoso server's SPARQL endpoint.

Service example: GetSuggestions

As an example for the LusTRE-ES functionalities and usage, consider the GetSuggestions operation. GetSuggestions is used to find concepts, which have a label starting with a given string (i.e. a sequence of characters typed in by the user). Case is ignored, *skos:prefLabel* and *skos:altLabel* are considered as labels. When specific languages are given as the input parameter, only labels in these languages are considered. Similarly, only specified vocabularies or service modules are used. By default, all available languages, vocabularies and service modules are used. The GetSuggestions operation supports the parameters listed in Table 4.

Two parameters can be used to control the extent to which the conceptual space spanned by the linked vocabularies is really explored in the query evaluation (for instance, when looking for synonyms, broader, narrower, or related concepts; or when looking for concepts represented by a given term or a term starting with the given character sequence):

- A "vocabularies" parameter enables the user to focus only on a subset of the available vocabularies when evaluating a request.
- The cross-walking parameter allows users to decide whether thesaurus interlinking should really be used when evaluating a request (instead of restricting the search to concepts all contained in one specific vocabulary). With cross-walking=true, LusTRE-ES considers all concepts linked by *skos:exactMatch* as the same concept. Specifying the cross-walking parameter can enhance GetSuggestions. With cross-walking=false, this generally leads to a smaller but more focused result set.

Web interface and the LusTRE-WEBe exploration tool

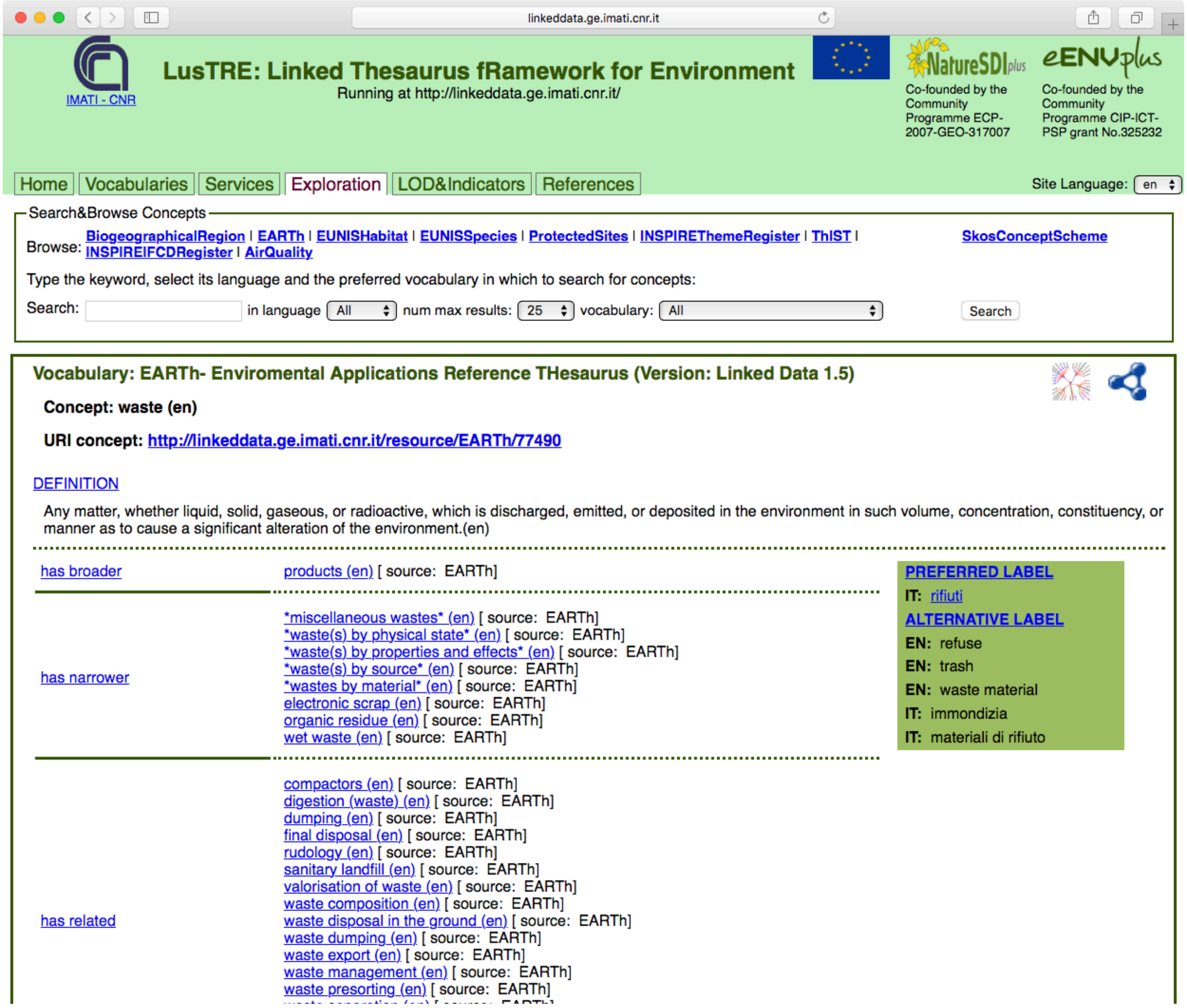
LusTRE is accessible at http://linkeddata.ge.imati.cnr.it. Users can navigate the following tabs: (i) LusTRE-VOC (by *vocabularies tab)* where all the information on the thesauri and code lists included in LusTRE (e.g. versions, licenses and provenance) are available; (ii) LusTRE-ES (by *service tab)* which provides the concise technical documentation required by third parties who want to exploit the web services in their clients; (iii) LusTRE-WEBe exploration tool (by *exploration tab)*, to provide a user-friendly client access to the knowledge infrastructure for browsing and searching for concepts in the vocabularies. In addition, the exploration tool also provides a visual browsing LusTRE-WEBeVis. It provides a graphical representation of the knowledge base extracted from the LusTRE-VOC through the visualization operation of LusTRE-ES. The visualization is activated from LusTRE-WEBe browsing after a concept has been selected.

The design and development of a user-friendly access to LusTRE-WEBe is a personalized Linked Data front-end to the Virtuoso SPARQL endpoint of LusTRE.

LusTRE-WEBe is able to:

- perform content negotiation compliant with Linked Data requirements (LOD Cloud 2014 requirements) and W3C best practice.
- support human readable browsing of the content. It clearly displays labels for the concepts, the definition of any relationships to other concepts (i.e. both internal to the thesaurus and those linked to other linked thesauri).

The LusTRE-WEBe tool provides two exploration practices, which exploit cross-walking to support query refinement and translations of concepts.



linkeddata.ge.imati.cnr.it

IMATI - CNR

LusTRE: Linked Thesaurus fRamework for Environment

Running at http://linkeddata.ge.imati.cnr.it/

NatureSDIplus: Co-founded by the Community Programme ECP-2007-GEO-317007

eENVplus: Co-founded by the Community Programme CIP-ICT-PSP grant No.325232

Home | Vocabularies | Services | Exploration | LOD&Indicators | References

Site Language: en

Search&Browse Concepts

Browse: BiogeographicalRegion | EARTh | EUNISHabitat | EUNISSpecies | ProtectedSites | INSPIREThemeRegister | ThIST | INSPIREIFCDRegister | AirQuality SkosConceptScheme

Type the keyword, select its language and the preferred vocabulary in which to search for concepts:

Search: in language All num max results: 25 vocabulary: All Search

Vocabulary: EARTh- Enviromental Applications Reference THesaurus (Version: Linked Data 1.5)

Concept: waste (en)

URI concept: http://linkeddata.ge.imati.cnr.it/resource/EARTh/77490

DEFINITION

Any matter, whether liquid, solid, gaseous, or radioactive, which is discharged, emitted, or deposited in the environment in such volume, concentration, constituency, or manner as to cause a significant alteration of the environment.(en)

has broader	products (en) [source: EARTh]
has narrower	*miscellaneous wastes* (en) [source: EARTh] *waste(s) by physical state* (en) [source: EARTh] *waste(s) by properties and effects* (en) [source: EARTh] *waste(s) by source* (en) [source: EARTh] *wastes by material* (en) [source: EARTh] electronic scrap (en) [source: EARTh] organic residue (en) [source: EARTh] wet waste (en) [source: EARTh]
has related	compactors (en) [source: EARTh] digestion (waste) (en) [source: EARTh] dumping (en) [source: EARTh] final disposal (en) [source: EARTh] rudology (en) [source: EARTh] sanitary landfill (en) [source: EARTh] valorisation of waste (en) [source: EARTh] waste composition (en) [source: EARTh] waste disposal in the ground (en) [source: EARTh] waste dumping (en) [source: EARTh] waste export (en) [source: EARTh] waste management (en) [source: EARTh] waste presorting (en) [source: EARTh] waste generation (en) [source: EARTh]

PREFERRED LABEL

IT: rifiuti

ALTERNATIVE LABEL

EN: refuse

EN: trash

EN: waste material

IT: immondizia

IT: materiali di rifiuto

Fig. 6 Example of search for the concept matching "waste"

Cross-walking for Query refinement. LusTRE-WEBe provides a Search and Browse entry point, which is useful when selecting the keywords for data discovery. By typing the initial letters of a desired keyword, the user can identify the corresponding concept and exploit the semantic relations of such a concept, making them coarser or finer in the initial search. The same techniques can also be used to discover more about concepts and terminologies pertaining to a specific domain. In addition, by exploiting cross-walking users can extend the browsing space in order to refine their query. The following is an example of query refinement: starting from the concepts in a thesaurus (i.e. "waste" in EARTh), the user can find additional related information in another thesaurus (e.g., "domestic waste landfill" in GEMET) exploiting the *skos:exactMatch* between "waste" in the two thesauri.

User start their search for a keyword, indicating a language and whether they are interested in searching the whole framework or only in one specific vocabulary, (e.g., type "waste" that is a keyword in English as the starting search concept). The tool returns all the concepts matching the keyword "waste" and the respective vocabulary source. By clicking on a concept (e.g., "waste" in EARTh), the user gets all the information related to it.

Figure 6 shows the concept description page for "waste" in EARTh. This page includes detailed information on: (i) the URI (http://linkeddata.ge.imati.cnr.it/resource/EARTh/77490) of the concept, which can be used by linked data applications to get the machine interpretable serialization of the concept; (ii) the lexical information of the concept, such as its definition, translations (e.g., the preferred label in Italian is "rifiuti") and synonyms (e.g., alternative labels in English such as "trash", "refuse", "waste material" and its alternative labels in Italian "immondizia" and "materiali di rifiuto"); (iii) the semantic relations with other concepts such as: broader concepts (e.g., "products"), narrower concepts (e.g., "electronic scrap", "organic residue") and related concepts (e.g., "dumping", "compactors").

Semantic relations enable users to expand their search. For example, by exploring broader terms, users can learn what the broader concepts mean and discover how they are related to the original concept. Users can continue their search by alternating textual and visual explorations. In order to browse texts, users click on web links that correspond to the semantically related concepts or they can activate node visualization for the visual navigation. For example, users could perform the following steps: (i) by text browsing they select the "waste" related concept "dumping", (ii) by visual browsing they move to "dumping" in GEMET linked by a *skos:exactMatch* from EARTh and select its related concept "landfill" in GEMET (Fig. 7); (iv) by continuing the browsing of GEMET, they get the narrow concept - "domestic waste landfill".

Figure 7 shows how to refine the query by cross-walking. Considering the concept "dumping" in EARTh, the visualization provides detailed information on dumping and its relation to other concepts within the same thesaurus or with the linked one. A compact view of the concept's structure is shown. Different node colours of the surrounding concept refer to the concepts of the different thesauri, while different edge colours refer to the different semantic relations (illustrated in the legend on the left).For example, the concept "dumping" has a *skos:exactMatch* with EuroVoc and with GEMET. Thus by cross-walking one of the two linksets, it is possible to navigate concepts from the two linked thesauri. Users can refine the query in GEMET: (i) Selecting only *narrowMatch* and *closeMatch* relations, the visualization shows only concepts with these kinds of matching. Selecting the concept "domestic waste landfill" in GEMET, a user moves from the concept "landfill" to its narrower concept "domestic waste landfill" in GEMET.

Cross-walking for translation. LusTRE-WEBe provides an entry point for searching the translations of a concept available in LusTRE (see Translation form in the exploration tab). It allows the user to select a concept belonging to a thesaurus in order to look for its alternative label in other languages within a thesaurus or by applying cross walking and navigating the matching concepts in the linked thesauri. Through cross-walking, translations that are not available for some languages in a vocabulary can be imported from equivalent concepts in other interlinked vocabularies.

For example, let us use the translation services without and with the cross-walking functionality to show the potential behind the interlinking. Let us suppose the user is interested in translating the word "soil" from the specific vocabulary EARTh which provides only Italian and English translations. First, the tool provides the set of concepts that match with "soil", and the user selects the concept to be translated. Without cross-walking, the tool provides synonyms available in EARTh that are only in English and Italian. When repeating the search by checking the cross-walking functionality, it returns not only translations and synonyms in EARTh but also all those in EARTh's interlinked thesauri. For example, it returns the translations available from AGROVOC (20 languages), GEMET (32 languages), and from the INSPIRE theme register (25 languages) whose intersection results in many more translations than EARTh alone.

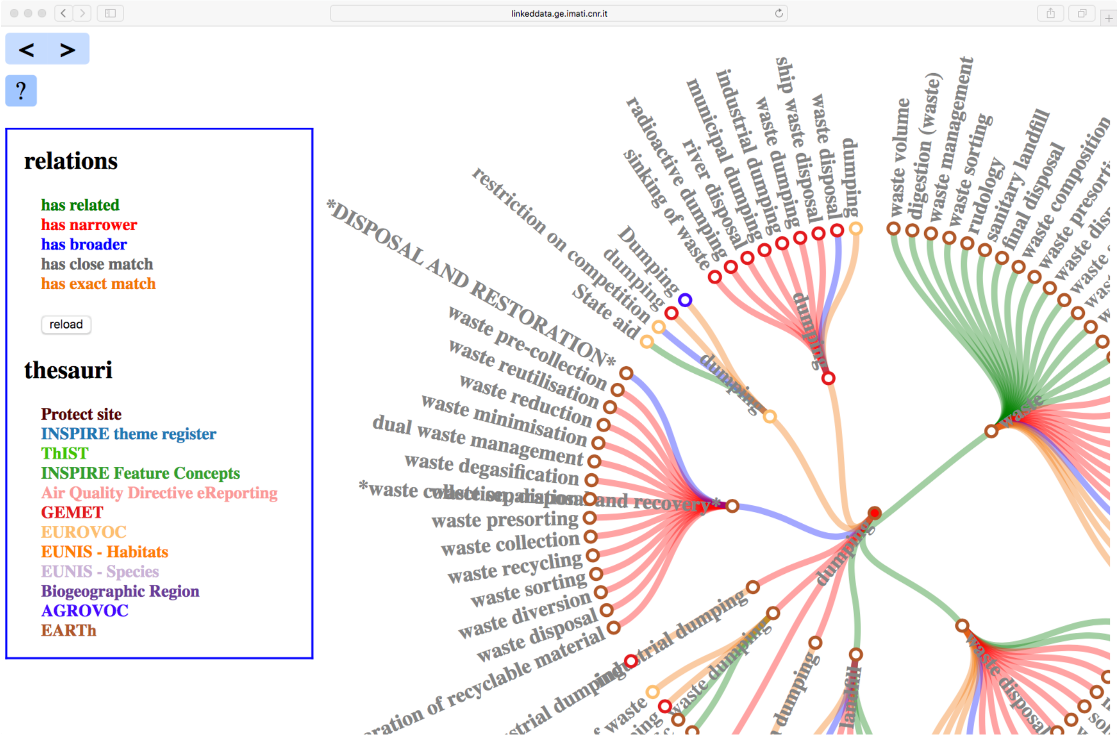


Fig. 7 Example of cross-walking: the conceptual neighbourhood of the concept "dumping" in EARTh and in GEMET

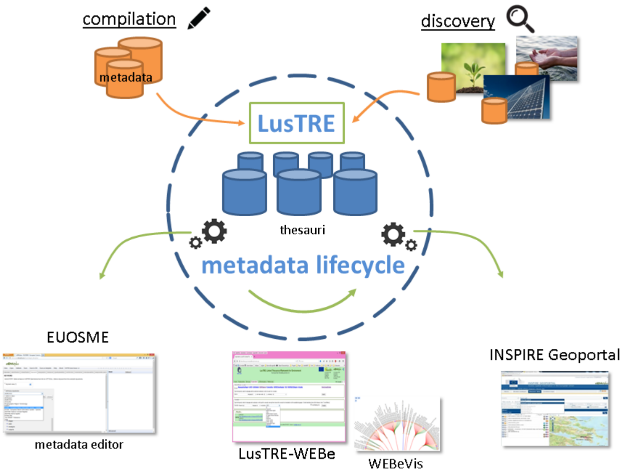


Fig. 8 Metadata lifecycle: from metadata compilation to data discovery.

LusTRE application for metadata management

The biggest value of LusTRE is in terms of the metadata management whenever third party tools access LusTRE (invisible/transparent for the user) through the LusTRE-ES and employ the LusTRE-VOC knowledge for improving and extending their own functionalities and services. Examples of such client applications are:

- Metadata editors that help users to write or edit metadata associated with environmental data sources or environmental information services within an SDI.
- Geodata portals or Catalogue Services for the Web, which are employed by a user for metadata-based finding and retrieving of environmental data or services.

Figure 8 shows how data provider/consumer can exploit LusTRE in the metadata lifecycle, from metadata compilation to data discovery. On the left, the figure shows the compilation task involving the data provider activity and tools involved (e.g., EUOSME metadata editor), and on the right, the task performed by the data-consumer and the tools involved (e.g., geoportal).

Users can use LusTRE in a separate web browser by searching for a concept in LusTRE-WEBe and then use this keyword in their own metadata editor. Alternatively, users can directly employ the EUOSME metadata editor, which prototypically exploits LusTRE services.

From a data-consumer perspective, thesauri and code lists facilitate data search and retrieval within a geoportal or when using a metadata catalogue. LusTRE can be deployed in the development of web services supporting resource selection.

For instance, the following features can be made available in a geoportal: (i) Auto-completion retrieves resources by providing hints of the keywords to be used according to the first letters typed by the user. (ii) Query expansion allows the retrieval of resources annotated with more specific concepts by exploiting narrower terms and also resources that are annotated with less specific concepts by exploiting the broader terms and related terms.

The metadata are used for common indexing. The concrete client components exemplarily extended are:

- LusTRE-WEBe tool to explore the vocabularies as a stand-alone tool.
- The European Open Source Metadata Editor EUOSME developed as part of EuroGEOSS project (http://www.eurogeoss.eu/).
- The developer version of the INSPIRE Geoportal (http://showcase.eenvplus.eu/client/geoportal.htm).

LusTRE-ES services are employed within the client application in order to exploit the linked vocabularies.

Table 5 summarizes the services that are employed in the aforementioned clients and their scope. The subsections below describe how the EUOSME editor and the INSPIRE Geoportal have exploited the LusTRE-ES services.

LusTRE exploitation in the EUOSME editor

The EUOSME metadata editor has been extended to navigate through the thesaurus framework by exploiting LusTRE-ES including extended features such as automated keyword completion, semantic search on keywords concepts, synonyms, and multilingual support (http://showcase.eenvplus.eu/client/editor.htm).

EUOSME automatically gets the list of available thesauri using GetCapabilities. The user can choose to navigate keywords from a thesaurus by selecting one from the list, or from all by selecting "All thesauri". Selecting the source of the keywords (from the list of thesauri), EUOSME connects to the server by asking for the list of the topmost concepts. The user may already enrich the metadata with one of these keywords or further refine, both navigating between relatives and searching by a string. Users can navigate between relatives using a tree-structure: by selecting a keyword GetRelatives uses the keyword chosen as a parameter, and this returns a list of related keywords.

Table 5 LusTRE exploitation services employed for tool extensions

TFES service	Used in	Supports users in
GetSuggestions	LusTRE-WEBe	Searching, browsing and translation.
GetCapabilities	LusTRE-WEBe	Formulating queries for searching and browsing.
GetSynonyms	LusTRE-WEBe	Getting richer results list in searching, browsing and translation.
GetTopMostConcepts	LusTRE-WEBe	Starting textual searching & browsing from the top concepts.
Visualization	LusTRE-WEBe	Visual TF navigation.
GetSuggestions	Geoportal	Rapid and correct formulation of queries.
GetRelatives	Geoportal	Most appropriate (re-)formulation of queries.
DescribeConcept	Geoportal	Understanding of unknown, difficult or ambiguous metadata terms.
GetSynonyms	Geoportal	Expanding query criteria achieving richer result sets.
GetSuggestions	EUOSME	More rapid and consistent metadata creation.
GetSynonyms, GetRelatives	EUOSME	Searches best fitting and most precise concepts in metadata compilation
GetTranslation	EUOSME	Multilingual metadata searches.
GetTopMostConcepts	EUOSME	Textual searching & browsing from the top concepts.

The free-text search uses GetSuggestions, which sends the search string and combines it with the chosen vocabularies.

If the user does not chose a language, the search is done in English. If no results in English are found, the results are returned in the first language available in the chosen vocabulary. If the user defines a language and does not find any results, the system automatically switches to English and so on. When a keyword is added to the metadata, the information on the source thesaurus (name, URL, publication date) and the URI of the keyword are also included.

LusTRE integration in the INSPIRE Geoportal

The INSPIRE Geoportal developed by Planetek Italian company for the European Joint Research Centre as the central portal for environmental information in the European Union was considered as a showcase of how a client can make use of the web services. A special version of the INSPIRE geoportal was thus developed (http://showcase.eenvplus.eu/client/geoportal.htm), which integrates the LusTRE knowledge base by exploiting the LusTRE-ES services in order to enhance data discovery. Initially, the INSPIRE geoportal used the GEMET thesaurus and INSPIRE Register code lists in order to provide suggestions that appeared when users started typing keywords in the search textbox of the "Discover Section" of the portal. To exploit LusTRE, the geoportal was decoupled from GEMET and was linked instead to the LusTRE thesaurus framework.

The client performs automatic multi-lingual keyword suggestions in the GUI (interactive discovery and auto-completion) given by LusTRE vocabularies. The suggestions are based on GetSuggestions and GetRelatives operations, the responses of which were integrated into the dropdown list of keyword suggestions that appear as soon as a seeker starts entering his search term.

The existence of multiple thesauri besides GEMET, significantly enriches user experience during query formulation and users can choose whether to exploit one or all the thesauri to retrieve suggestions and relatives.

A third operation, the DescribeConcept was added to helps users in understanding obscure or ambiguous metadata terms, thus bridging a frequent familiarity gap. This problem is resolved by displaying tooltips with a concept definition when users hover the cursor over that specific keyword. DescribeConcept is predicated on the existence of a definition for a particular concept within the thesauri being queried.

Another integrated operation is GetSynonyms. Unlike the first three, which are used at the user-interface during query formulation, this operation works in the background and is aimed at a better and more inclusive retrieval of results. Specifically, after the selection of a keyword to query the system with, but before submitting the search term to the backend, a GetSynonyms operation is performed to retrieve other terms that are semantic synonyms for the original search term. The query is expanded with these additional terms in order to retrieve more resources that will satisfy the user's query. This is particularly useful in discovering results in languages other than the one in which the query term is expressed.

All four operations were integrated by adding new code to the geoportal that exploits the interface of LusTRE-ES. The Geoportal formulates the requests through LusTRE-ES services, parses the services' responses to extract the necessary results and shows results on the user interface.

LusTRE access

LusTRE is accessible at http://linkeddata.ge.imati.cnr.it.

LusTRE-VOC can be accessed in:

- HTML View: Users can browse LusTRE-VOC using the Tab Exploration, which provides access to the Exploration Tool LusTRE-WEBe component and to the access point for Semantic Explorative Search by the Visual Exploration Tool LusTRE-WEBeVis;

- RDF View: Users can explore the framework with semantic web browsers such as Tabulator or Disco, starting browsing from top most concepts;
- SPARQL Endpoint: SPARQL clients can query the framework at this SPARQL endpoint: http://linkeddata.ge.imati.cnr.it:8890/sparql.

Exploitation by a third party is performed through the web services whose description is documented at http://linkeddata.ge.imati.cnr.it/services.jsp.

The use of the LusTRE component is free. In particular: (i) thesauri and code lists included in LusTRE are licenced as specified in LusTRE web page; (ii) all the interlinkings are released under the BY-CC licence; (iii) The service code developed is released under the Apache Licence 2.0 and is available at https://github.com/eENVplus/tf-exploitation-server.

For inquiries on how to re-use this software: lustre@ge.imati.cnr.it.

Discussion and conclusions

We have described a multilingual thesaurus framework developed within the EU-funded project eENVplus to tackle cross-sectorial and cross language issues in data sharing. We developed a framework of multilingual linked thesauri (also adding a few code lists), namely LusTRE, in order to improve environmental data interoperability. It provides a wide terminological content of thesauri and code lists available in the LusTRE-VOC knowledge infrastructure and a set of web services, LusTRE-ES, to exploit such content for metadata compilation and data discovery.

In the paper, we have outlined a technical multi-phase methodology for the implementation of a linked thesaurus framework employing the linked data technologies available at the time of the project. The outcome is the publication and/or mapping to existing relevant thesauri and also small size code lists used by cross-sectorial communities working for the Environment.

The adoption of the Linked Data technology led to the creation of an open, flexible and exploitable environment of thesauri and code lists. LusTRE also provides a set of web services for vocabulary exploitation. In particular, the cross-walking feature provides a full exploitation of the mapping between vocabularies, allowing users to navigate among matching concepts in a transparent way.

We also show how to effectively exploit LusTRE through two different approaches: (i) as a stand-alone tool, thanks to the availability of an exploration tool, which we developed during the project, allowing both textual and visual browsing of the vocabularies; (ii) integrated through the web services in client applications such as metadata editor and geoportal.

We exploited LusTRE in the metadata editor EUOSME and the INSPIRE Geoportal employed in ten pilot studies developed within the eENVplus project. All the pilots involved populating a metadata catalogue related to various INSPIRE data themes so that all the applications and the eENVplus infrastructure would be able to harvest data. Each pilot exploited LusTRE vocabulary to enrich the metadata file with several keywords retrieved from the set of shared vocabularies. The goal was to enhance the level of interoperability between datasets providing value for metadata elements with a large and cross-sectorial and multi-lingual terminology provided by the linked vocabularies. By exploiting LusTRE, users can describe data in a uniform way by selecting keywords from existing shared code lists and thesauri, avoiding as much as possible the use of free text keywords. A uniform data description increases the interoperability: through the exploitation of LusTRE-ES by the catalogue client, it is possible to discover data searching by keywords, synonyms, narrower concepts, etc.

We believe that our work can contribute to a wider adoption of the INSPIRE initiative by providing the Environment communities with a wide terminology for sharing their data by providing a uniform data description and improving data indexing and searches among the different communities.

The thesaurus framework helps: (i) data providers with a faster, more consistent and better informed metadata editing; (ii) data providers (and INSPIRE data providers) to enter a new domain more easily and faster and to bridge between different domains, disciplines and languages: (iii) data consumers to find more precise, more comprehensive, cross-lingual and cross-domain resources; (iv) the public and newcomers in a certain domain to find the knowledge structures in this domain faster, to build bridges into other domains, and to span across disciplines and languages.

In terms of tool providers and infrastructure architects: (i) LusTRE provides the opportunity for developers of domain-specific thesauri to easily publish their terminologies with novel and powerful technologies (i.e., LOD). Publishing into the LOD makes terminologies easily exploitable and inter-linkable such that the value of each terminology will be leveraged through the integration into other tools and through the combination with other researchers' works. Publishers need to contact the LusTRE team for publishing new terminologies in LusTRE. (ii) For developers of INSPIRE-implementation tools (for metadata management, for data searches, for CSW, etc.), the thesaurus framework approach provides an opportunity to access environmental thesauri through simple and stable service interfaces without having to maintain the thesauri themselves or creating software for their exploitation. (iii) For developers of thesaurus-exploitation services, LusTRE offers a modular and easily extensible architecture for integrating ideas and service modules into a larger, established framework, thus leading to much more visibility and impact.

Web accesses to LusTRE have been monitored through the Tomcat Log. The log results from 2017 show that there were almost 2,000,000 accesses to LusTRE, and 34,672 accesses to services. Currently LusTRE is also being exploited by QSphere plugin (https://www.fgdc.gov/iso-metadata-editors-registry/editors/11).

LusTRE acknowledges the importance of tracking the quality of published KOS and linksets. Domain experts have double-checked the correctness of the discovered links, and in-house quality metrics have been developed and applied to assess the value of links in terms of browsability and multilingual gains (Albertoni et al. 2016a, 2016b). The quality of the thesauri is assessed by qSKOS (Mader 2012). We have extended the qSKOS open software, encoding its results according to the recent W3C Data Quality Vocabulary (Albertoni et al. 2016c) suggested by the Data on the Web Best practices (Calegari et al. 2017). The results of the LusTRE quality analysis are uploaded into the LusTRE SPARQL endpoint in order to promote a conscious use of the LusTRE resources.

LusTRE contributes in the joint exploitation of the available KOSs. It federates the existing KOSs by interlinking them and providing a unique selling point. However, LusTRE has its deficiencies. The provision of interlinking is a loosely coupled solution that represents a trade-off between effort and benefits. A deeper level of consistency and coherence between heterogeneous KOSs is beyond the scope of the activity illustrated in this paper, but might be required for some specific applications. Inter-KOSs editorial committees and collaborative tools would ease the maintenance and improve the harmonization of federated KOSs. Knowledge changes continuously linking between vocabularies may become inconsistent. Versioning and regular consistency of links between vocabulary checks need to be addressed by the LusTRE curator. We plan to fully adopt the recent W3C Data on the Web Best Practices recommendation (Calegari et al. 2017). This should promote LusTRE to a larger Semantic Web community as well as to the spatial data communities interested in improving their technology with respect to W3C standard recommendations.

Acknowledgements. This work has been supported by the European Commission through the project eENVplus (grant no. CIP-ICT-PSP 325232, http://www.eenvplus.eu/)

References

Albertoni R, De Martino M, Di Franco S, De Santis V, Plini P (2014a) EARTh: An Environmental Application Reference Thesaurus in The Linked Open Data Cloud. In Semantic Web, IOS Press 5(2): 165-171

Albertoni R, De Martino M, Podestà P (2014b) Environmental thesauri under the lens of reusability. In: Kő A., Francesconi E. (eds) Electronic Government and the Information Systems Perspective. EGOVIS 2014. Lecture Notes in Computer Science, vol 8650. Springer, Cham, pp 222-236

Albertoni R, De Martino M, Podestà P (2016a) Linkset Quality Assessment for the thesaurus framework LusTRE. In: Garoufallou E, Subirats Coll I, Stellato A, Greenberg J. (eds) Metadata and Semantics Research. MTSR 2016. Communications in Computer and Information Science, vol 672. Springer, Cham, pp 27-39

Albertoni R, De Martino M, Quarati A (2016b) Integrated Quality Assessment of Linked Thesauri for the Environment. In: Kő A, Francesconi E. (eds) Electronic Government and the Information Systems Perspective. EGOVIS 2016. Lecture Notes in Computer Science, vol 9831. Springer, Cham: 221-235

Albertoni R, Isaac A, Debattista J, Dekkers M, Guret C, Lee D, Mihindukulasooriya N, Zaveri A (2016c) Data on the Web Best practices: Data Quality Vocabulary, W3C Working Group Note. http://www.w3.org/TR/vocab-dqv/

Attardo C, Saio G (2013) eENVplus: A Framework To Support eEnvironmental Services And Applications. In 27th Int. Conf. on Environmental Informatics for Environmental Protection, Sustainable Development and Risk Management, EnviroInfo-2013, B. Page et al, eds., pp 684-692

Baker T, Bechhofer S, Isaac A, Miles A, Schreiber G., Summers E (2013) Key choices in the design of Simple Knowledge Organization System (SKOS). J. Web Sem 20: 35-49

Baker T, Caracciolo C, Doroszenko A, Suominen O (2016) GACS Core: Creation of a Global Agricultural Concept Scheme. In: Garoufallou E., Subirats Coll I., Stellato A., Greenberg J. (eds) Metadata and Semantics Research. MTSR 2016. Communications in Computer and Information Science, vol 672. Springer, Cham, 311-316

Berners-Lee T (2009) Linked Data. http://www.w3.org/DesignIssues/LinkedData.html. Last access: 29 March 2016

Caracciolo C, Stellato A, Morshed A, Johannsen G, Rajbhandari S, Jaques Y, Keizer J (2013) The AGROVOC Linked Dataset. In: Semantic Web 4(3): 341-348

Carusone A, Olivetta L (2006) Italian Thesaurus of Earth Sciences, APAT. http://www.isprambiente.gov.it/contentfiles/00003600/3687-thist.pdf

Calegari N, Burle C, Bernadette Farias L, Greiner A, Isaac A, Iglesias C, Laufer C, Guéret C, Lee D, Schepers D, Stephan E G, Kauz E, Atemezing G A, Beeman H, Bittencourt I, Almeida J P, Makx Dekkers M,Winstanley P, Archer P, Albertoni R, Purohit S, Córdova Y (2017) Data on the Web Best Practices, W3C Recommendation. https://www.w3.org/TR/2017/REC-dwbp-20170131/, Last access: 1 July 2017

Da Silva OC, Lisboa-Filho J, Braga JL et al. (2009) Searching for metadata using knowledge bases and topic maps in Spatial Data Infrastructures, Earth Science Informatics 2 (4): 235–247

De Martino M, Albertoni R (2011) A Multilingual / Multicultural Semantic-Based Approach To Improve Data Sharing in an SDI for Nature Conservation. In: Int. Journal of Spatial Data Infrastructures Research 6: 206-233

EIONET (2015) GEMET Thesaurus. http://www.eionet.europa.eu/gemet

GS SOIL (2015) GS SOIL project. https://inspire.ec.europa.eu/SDICS/gs-soil

Hyland B, Atemezing G, Villazón-Terrazas B (2016) W3C Working Group Note: Best Practices For Publishing Linked Data, http://www.w3.org/TR/ld-bp/. Last Access: 29 March 2016.

Heath T, Bizer C (2011) Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, 1:1

INSPIRE (2015a) Infrastructure for Spatial Information in the European Community. http://inspire.ec.europa.eu/

INSPIRE Implementing Rules (2015b) http://inspire.ec.europa.eu/inspire-implementing-rules/51763

INSPIRE (2015c) INSPIRE registry http://inspire.ec.europa.eu/registry

ISO 19115-1 (2014) Geographic information--Metadata. https://www.iso.org/standard/53798.html

L'Abate G, Caracciolo C, Pesce V, Geser G, Protonotarios V, Costantini E (2015) Exposing vocabularies for soil as Linked

Open Data, Information Processing in Agriculture 2 (3-4): 208-216

Mader C, Haslhofer B, Isaac A (2012) Finding Quality Issues in SKOS Vocabularies. In: Zaphiris P, Buchanan G, Rasmussen E, Loizides F (eds) Theory and Practice of Digital Libraries. TPDL 2012. Lecture Notes in Computer Science, vol 7489. Springer, Berlin, Heidelberg, pp 222-233

Manaf N A A, Bechhofer S, Stevens R (2012) The Current State of SKOS Vocabularies on the Web. In: Simperl E, Cimiano P, Polleres A, Corcho O, Presutti V. (eds) The Semantic Web: Research and Applications. ESWC 2012. Lecture Notes in Computer Science, vol 7295. Springer, Berlin, Heidelberg, pp 270-284

Mazzocchi F, Plini P (2005) Development of the Environmental Applications Reference Thesaurus (EARTh). Proceedings of the 7th ISKO-Spain Conference, The human dimension of knowledge organization, Barcelona, Spain, Jul. 6-8

Miles A, Bechhofer S (2009) W3C Recommendation: Simple Knowledge Organization System Reference. http://www.w3.org/TR/skos-reference. Last Access: 20 March 2014

Moué P, Ortmann J (2009) Getting across information communities. Embedding semantics in the SDI for the Amazon. Earth Sci Inform 2: 217-233

Palavitsinis N, Manouselis N (2009) A Survey of Knowledge Organization Systems in Environmental Sciences. In I. Athanasiadis, A. Rizzoli, P. Mitkas, and J. Gomez, editors, Information Technologies in Environmental Engineering, Springer Berlin Heidelberg, pp 505-517

Schandl T, Blumauer A (2010) PoolParty: SKOS thesaurus management utilizing linked data. In: Aroyo, L., Antoniou, G., Hyvönen, E., Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.). The Semantic Web: Research and Applications 6089, Springer, Heidelberg, pp 421–425

Schreiber G, Amin A, Aroyo L (2008) Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. Web Semantics: Science, Services and Agents on the World Wide Web 6 (4): 243- 249

Stellato A, Rajbhandari S, Turbati A, Fiorelli M, Caracciolo C, Lorenzetti T, Keizer J, Pazienza MT (2015) VocBench: A Web Application for Collaborative Development of Multilingual Thesauri. In: Gandon F, Sabou M, Sack H, d'Amato C, Cudré-Mauroux P, Zimmermann A (eds) The Semantic Web. Latest Advances and New Domains. ESWC 2015. Lecture Notes in Computer Science, vol 9088. Springer, Cham, pp 38-53

Stellato A, Turbati A, Fiorelli M, Lorenzetti T, Costetchi E, Laaboudi C, Van Gemert W, Keizer J (2017) Towards VocBench 3: pushing collaborative development of thesauri and ontologies further beyond, Proc. of the 17th European Networked Knowledge Organization Systems Workshop (NKOS 2017). CEUR Workshop Proceedings

Wolstencroft K, Owen S, du Preez F, Krebs O, Mueller W, Goble C (2011) The SEEK: a platform for sharing data and models in systems biology. Methods Enzymol. 2011; 500:629–55 doi: 10.1016/B978-0-12-385118-5.00029-3

Wright D, Harrison K, Watkins J (2015) Automated tagging of environmental data using a novel SKOS formatted environmental thesaurus, Earth Sci. Inform 8(1): 103–110